APPENDIX A. STATISTICAL SUPPORT FOR SAMPLING DESIGNS

Table of Contents

Tables

Figures

Acronyms

95UCL	95% upper confidence limit for the mean
ARAR	applicable or relevant and appropriate requirement
AWQC	ambient water quality criteria
cfs	cubic feet per second
COC	contaminant of concern
cPAH	carcinogenic polycyclic aromatic hydrocarbon
CLT	Central Limit Theorem
CV	coefficient of variation
DL	detection limit
dw	dry weight
EAA	early action area
ENR	enhanced natural recovery
FS	feasibility study
GOF	goodness-of-fit
LDW	Lower Duwamish Waterway
MDD	minimum detectable difference
MNR	monitored natural recovery
NTR	National Toxics Rule
PCB	polychlorinated biphenyl
PE	polyethylene
ppb	parts per billion
PSS	practical salinity scale
RAO	remedial action objective
RI	remedial investigation
RM	river mile
RME	relative margin of error
ROD	Record of Decision
SD	standard deviation

SE	standard error
SWAC	spatially weighted average concentration
TEQ	toxic equivalent
TSS	total suspended solids
TTL	target tissue level
USGS	US Geological Survey
WAC	Washington Administrative Code
WQC	water quality criteria
ww	wet weight

1 Introduction

This appendix presents information relevant to ~~A statistical evaluation was conducted to support the~~ sampling designs and data analysis of ~~for the collection of~~ surface sediment (0–10 cm), intertidal sediment (0–45 cm), ~~and~~ fish and crab tissue, clam tissue, and surface water.

For each of the sampled media (except surface water), methods to calculate 95% upper confidence limits for the mean (95UCLs), to compare to cleanup levels or target tissue levels (TTLs), are presented. -Formulas for each sampled medium are provided based on the current expectation of the statistical distribution for the data. Once collected, each dataset will be evaluated and the most appropriate method for calculating the 95 UCL will be used.

This appendix is organized by media type:

- Section 2 ~~summarizes~~ presents ~~the data that were used to estimate post-remediation variances, which are used to plot the relationship between sample size and RME.~~ statistical evaluation for surface sediment (0–10 cm) sampling design. ~~The data from surface sediments (0–10 cm) and fish and crab tissues were extracted from the RI sediment and tissue datasets (Windward 2010a).~~
- Section 3 presents ~~the~~ statistical evaluation for the intertidal sediment (0-45 cm) sampling design~~presents the methods used to make meaningful estimates from the previous datasets to predict patterns of variability for baseline and future monitoring, and to develop optimal sampling designs. Sampling methods, sampling frames (i.e., areas targeted for sampling), and sampling objectives are some of the ways in which the RI datasets differed substantially from baseline and future LDW sampling efforts. The statistical approach used in this appendix to extract summary statistics relevant to the development of the baseline sampling design are detailed for surface sediments (Section 3.1) and fish and crab tissues (Section 3.2).~~
- Section 4 presents ~~the results from the methods outlined in Section 3, including a summary of the distributional characteristics and estimates of variability for sediments (Section 4.1) and fish and crab tissues (Section 4.2).~~statistical evaluation for the fish and crab tissue sampling design
- Section 5 presents statistical evaluation for the clam tissue sampling design.~~presents methods for calculating the compliance metric (95UCL) for surface sediments and fish and crab tissues, intertidal sediments (0–45 cm) for direct contact during beach play and clamming scenarios, and clam tissues.~~
- Section 6 presents statistical evaluation for the surface water sampling design.~~discusses the conclusions regarding sampling variability for the study~~

~~designs considered herein, and makes final recommendations for sediment and tissue baseline and future monitoring.~~

- Section 7 presents the references.

2 Surface Sediment (0–10 cm)

To develop the surface sediment sampling design, data from monitored natural recovery (MNR) areas identified in Record of Decision (ROD) Figure 18[1] (EPA 2014) were used to estimate the magnitude and patterns of variability expected in the Lower Duwamish Waterway (LDW) following active remediation.[2]

The targeted relative margin of error (RME) for the site-wide mean concentration for surface sediments (0–10 cm) in the LDW is ≤ 25%, wherein the RME is calculated as the width of the ~~95% upper confidence limit for the mean (~~95UCL~~)~~ as a percent of the mean. ~~[3]~~ The sampling objective to estimate the site-wide mean with a RME of 25% can be met most cost-effectively through the use of composite samples. Composite samples are~~will~~ not intended to provide information regarding the population variance of individual sediment chemical concentrations, nor details of small-scale spatial heterogeneity. That information will be collected through area-specific sampling during remedial design. The baseline surface sediment (0–10 cm) sampling design will provide an efficient estimate of the 95UCL of the site-wide mean ~~for~~ to compare to cleanup goals for remedial action objectives (RAOs) 1, 2, and 4.[4]

A spatially balanced sampling design has been developed that includes the collection of a single random sample within *n* (e.g., *n* = 100, 140, 150, or more) ~~each of 100~~ grid cells of approximately equal area distributed throughout the LDW~~river~~. ~~Twenty~~ C~~c~~omposite samples with the same number of field samples in each are ~~then~~ constructed from groups of *k* neighboring individual samples for analysis. The sample size of analytical samples is *n/k* (e.g., 100 field samples would be used to create 20 composites with 5 samples each). This approach avoids bias and spatial clustering of samples so that the arithmetic mean of the observations is~~, in effect,~~ also a spatially weighted average concentration (SWAC), because equal spatial weighting is intrinsic to the sample design.

[1] ROD Figure 18 is titled Selected remedy.

[2] It is acknowledged that baseline sediment chemistry variability will likely be greater than the variability estimated from the MNR dataset, and may be skewed rather than symmetric (i.e., follow a gamma distribution rather than a normal distribution).

~~[3] See Section 5.1 for more details.~~

[4] RAO 1 is related to consumption of resident seafood (human health), RAO 4 is related to high-trophic-level ecological risks (river otter), and RAO 2 is related to direct contact (human health) from netfishing (using 0–10-cm sediment samples throughout the LDW) and clamming and beach play (using 0–45-cm sediment sampling in specified areas).

In future years of site-wide monitoring for RAOs 1, 2, and 4, the number of samples per composite should remain consistent to maintain year-to-year comparability of the datasets. The numbers of field samples and composite samples may change in response to updated information about site variance, and to achieve a desired RME for the site-wide mean. ~~the boundaries of the sampling grid cells and the compositing scheme among grid cells will remain constant. The sample locations will be randomly placed within each grid cell for each monitoring event, providing an unbiased characterization of each grid cell at that point in time. With this sampling design, the baseline and each future survey will maintain the same connection to a spatial area within the site (e.g., composite sample No. 1 will always provide an unbiased estimate of the mean of the spatial area defined by grid cells 1 through 5).~~ In this way, a robust site-wide 95UCL can be calculated for each sampling event.

The site-wide results for baseline sediment sampling will be used to chart the progression of sediment concentrations toward the cleanup goals. When sufficient sampling events have been completed (e.g., five or more), the trend for these data can be estimated using regression or correlation methods. In the interim, the baseline data set may be used most simply in a two-sample, one-tailed comparison to a data-set collected in one of the future sampling events. The specific statistical test used will depend on the nature of the data-sets (e.g., distribution, equality of variance, number of non-detects). When non-detects are present, Kaplan-Meier estimates of mean and variance will be used, as well as substitution at full detection limit (DL) and at 0 to provide upper and lower bounds for population estimates.

Sections 2.1 through 2.3 of this appendix present analyses using existing data to illustrate the level of variability expected within the LDW following active remediation. These data support a sampling design with 20 composite samples from 140 field samples (5 samples per composite) to achieve the targeted RME of 25% or better during post-remedy sampling, 90% statistical power to detect a 60% decrease in the site-wide PCB SWAC,[5] and a sampling density within 1.5 times the minimum separation distance, on average. However, after reviewing these results and considering the age and spatial representation of the dataset on which they were based, EPA directed a more conservative assumption regarding variance, which resulted in a sampling design with 24 composite samples of 7 samples each (total of 168 field samples). The EPA-directed sampling design is presented in the Work Plan (Windward and Integral 2017), Section 3.2.1.1.

2.1 Surface Sediments (0–10 cm) Data Used in the Analysis

Surface sediment data from MNR areas (as depicted in ROD Figure 18 and Map ~~B~~A-1 of this appendix) within the RI/feasibility study (FS) dataset were used in this evaluation.

[5] The feasibility study (FS) estimated a decrease in the site-wide PCB SWAC of approximately 60% between post-EAA conditions (i.e., baseline) and post-remedy conditions (FS Table 9-2).

Data from MNR areas were selected because they provided the best surrogate~~s~~ for data variability likely to exist following active remediation in the LDW.[6] Results for total PCBs (sum of Aroclors]), carcinogenic polycyclic aromatic hydrocarbon (cPAH) toxic equivalent (TEQ), and arsenic were evaluated.[7]

A summary of the data for each contaminant of concern (COC) by river mile (RM) segment is presented in Table A-1. The three COCs were mostly detected in this dataset (i.e., 88% of the PCB samples had detected concentrations and 95% of the cPAH and arsenic samples had detected concentrations). The data for total PCBs were the most abundant, with sample counts within each segment ranging from 8 to 103 for total PCBs and from 4 to 61 for both cPAH and arsenic. Sample locations within segments were clustered to varying degrees throughout the site; nearest neighbors were less than 50 ft apart in all but one segment for total PCBs, and in all but four segments for cPAH and arsenic.

[6] The only data that were excluded from the MNR area dataset were perimeter samples associated with early action areas (EAAs) (Terminal 117, Slip 4, and Duwamish Diagonal) that were collected prior to remediation and had polychlorinated biphenyl (PCB) concentrations greater than 400 µg/kg dry weight (dw).

[7] The sums of PCB Aroclors and cPAH TEQ were calculated using the LDW RI/FS data management rules. The dioxin/furan data are limited and thus no evaluation has been conducted for dioxin/furan TEQs.

Table A-1. Summary of surface sediment (0–10 cm) data from MNR areas within the RI/FS dataset used to evaluate statistical properties of proposed study designs

RM Segment [a]	Total PCBs (ug/kg, dw)				cPAH TEQ (ug/kg, dw)				Arsenic (mg/kg, dw)			
	Total N	No. NDs	Concentration Range[b]	Min. Distance Between Samples (ft)	Total N	No. NDs	Concentration Range	Min. Distance Between Samples-(ft)	Total N	No. NDs	Concentration Range	Min. Distance Between Samples (ft)
0,0.1	16	1	8.4–250	21	14	1	44–720	21	14	0	5.10–21.2	21
0.1,0.3	17	3	3.1–191	1	14	1	9.1–760	1	14	0	3.50–21.9	1
0.3,0.5	16	0	7.0–222	46	13	0	20–530	97	13	1	3.10–17.0	97
0.5,0.6	18	4	0.4–341	1	21	2	4.3–880	1	21	3	3.10–33.9	1
0.6,0.7	16	2	2.6–340	46	11	0	30–480	46	11	0	5.8,0 13.0	46
0.7,0.9	19	1	4–196	58	11	0	34–860	87	11	0	3.10–20.2	87
0.9,1	8	0	51–240	20	4	0	320–660	210	4	0	9.10–31.8	210
1,1.2	16	1	4–302	12	12	0	350–550	62	12	0	9.50–45.0	62
1.2,1.4	13	0	66–290	11	10	0	160–670	11	10	0	8.80–46.8	11
1.4,1.6	17	2	10–340	25	13	0	17–500	25	15	0	1.20–16.7	25
1.6,1.8	30	6	9.5–270	11	22	0	21–520	11	25	0	2.40–26.0	11
1.8,2	16	1	9.5–260	17	13	0	48–890	17	14	0	5.10–17.7	17
2,2.1	13	0	38–296	37	10	0	27–650	49	10	0	4.20–23.1	49
2.1,2.7	32	1	10–204	35	21	2	9.1–1,000	35	26	3	1.80–17.6	35
2.7,2.9	31	0	36–380	13	9	0	34–320	38	9	0	9.00–26.5	38
2.9,3.2	34	2	10–162	6	18	0	32–250	6	18	0	4.90–11.5	6
3.2,3.7	45	5	7.1–380	10	15	0	61– 320	35	18	5	6.50–13.6	35
3.7,4.1	37	4	0.4–370	16	26	0	9.7–210	25	24	0	4.80–14.4	25
4.1,4.6	103	12	3–340	6	61	2	9.4–1400	6	61	0	3.50–17.8	6
4.6,5	48	18	0.3–162	9	36	11	9.4–1060	18	34	3	1.90–51.0	18
Total	545	63			354	19			364	15		

[a] Square brackets are inclusive: [0, 0.1] indicates locations with 0 ≤ RM ≤ 0.1. Left parenthesis indicates strictly greater than: (0.1, 0.3] captures locations with 0.1 < RM ≤ 0.3.

~~[b] Concentration ranges are provided for PCBs because the CV for PCBs was determined to be the most accurate and thus was used to develop the sampling design.~~

cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
dw – dry weight
MNR – monitored natural recovery
ND – non-detect
PCB – polychlorinated biphenyl
RI/FS – remedial investigation/feasibility study
RM – river mile
TEQ – toxic equivalent

The highly clustered nature of the historical sampling locations (Map ~~B~~A-1 and minimum distance noted in Table A-1) made it inappropriate to calculate simple summaries of the mean and variance of the data within each segment. Instead, a simplified bootstrap estimate[8] of the coefficient of variation (CV) for each of the three COCs indicated that the site-wide CVs for total PCBs and cPAH TEQ were similar, while the CVs for arsenic were slightly lower (Table A-2). Using the highest CVs to inform the study design provides appropriate estimates of the expected RME for the most variable analytes, and a buffer on the expected RME for analytes with lower CVs. Although the CVs for total PCBs and cPAH TEQ were similar, the CV for total PCBs was considered more accurate because there were approximately 200 ~~many~~ more total PCB samples than cPAH TEQ samples ~~have been~~ analyzed throughout the LDW (Table A-1). Consequently, the remainder of the sediment discussion in this appendix presents results from only the total PCBs data; it is assumed that the study design based on PCB data will result in similar or better RME values for the other COCs.

Table A-2. Distribution of the surface sediment CVs across bootstrap replicates

COC	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Total PCBs	0.6	0.8	0.8	0.8	0.8	1.0
cPAH TEQ	0.7	0.8	0.8	0.8	0.9	1.0
Arsenic	0.3	0.5	0.5	0.5	0.6	0.7

Note: Each bootstrap replicate (B = 1,000) was comprised of 100 observations randomly selected from the RI surface sediment dataset (Map ~~B~~A-1), with the stipulation that all sampling locations were separated by at least 200 ft.

COC – contaminant of concern
cPAH – carcinogenic polycyclic aromatic hydrocarbon
CV – coefficient of variation
PCB – polychlorinated biphenyl
RI – remedial investigation
TEQ – toxic equivalent

2.2 Surface Sediments (0–10 cm) Methods

The sampling programs represented in the RI/FS dataset used a variety of sampling designs with different objectives, and many of the sampling programs focused on smaller areas. As a result, the RI/FS dataset has irregular sampling densities across the site, including some areas with very tightly clustered samples and other areas with very few samples (Table A-1, Map ~~B~~A-1). Using spatially clustered samples as if they were independent samples would likely result in biased estimates of mean and variance, which would not be representative of the expected site-wide conditions following active remediation.

Sampling variance is the variability of summary statistics (e.g., the mean) if the same sampling design, with the same sample size, were applied to the same population multiple times. A lower sampling variance results in improved precision in estimates of summary statistics. Some ways to reduce sampling variance include:

[8] Each bootstrap replicate (B = 1,000) drew a random sample of 100 observations without replacement from the RI dataset, with the stipulation that sampling locations were separated by at least 200 ft.

- Reducing variance among samples (e.g., by analyzing composites of multiple grab samples, thereby averaging over smaller scale variability)
- Increasing sample size throughout the site (e.g., a mean of 100 samples has lower variance than a mean of 20 samples)
- Using a stratified sampling design (e.g., by having higher sample densities within areas [strata] with higher variance or different means to reduce the sampling variability in the overall mean)

To use the existing data from MNR areas (Map ~~B~~A-1) to assess the benefits of different sampling approaches and determine which could be most efficiently used to improve precision, three key questions were asked. These questions, and the methods used to answer them, are described below.

2.2.1 Question 1

What minimum separation distance between samples would be required to produce spatially independent data?

The minimum separation distance between samples was required to reduce the bias and redundancy of information resulting from the tightly clustered samples within the RI dataset. The minimum separation distance was used to restrict how the data within the RI dataset were sampled in the bootstrapping exercise.

Method: A correlogram displays the average spatial correlation (Moran's I) between pairs of samples within increasing distance intervals. The distance interval at which the correlation becomes nominal was used to determine the minimum separation distance. Correlograms were created using two different functions in R (R Core Team 2016): *correlog{pgirmess}* (Giraudoux 2016) and *correlog{ncf}* (Bjornstad 2016).

2.2.2 Question 2

What is the variance of concentrations within different reaches of the LDW, and is it approximately consistent throughout the LDW?

If the spatial variance were very different within different sections of the river, this would indicate that variance strata exist and precision of the site-wide mean could be improved by stratifying the river and taking more samples where variance is higher.

Method: Random groups of five adjacent samples were bootstrapped from the RI dataset. A sample size of five was chosen to mimic the sample sizes that will be used in composite sampling, and 5,000 bootstrap samples were drawn. Within each group, the randomly selected samples were separated by the minimum distance established by the answer to Question 1, and no more than a maximum distance of 1,320 ft (0.25 mi). This maximum separation distance was used because it was large enough to not limit the number of bootstrapped sample groups that could be formed, but not so large as to average over spatial patterns in concentrations that were present in this dataset. The variability within these groups of five samples was plotted against location along the river (average river mile of the five samples within the group). Any large changes in the magnitude of variance at different river miles would support the use of a stratified sampling design.

2.2.3 Question 3

What is the expected sampling variance for the LDW-wide mean using a set of 100 spatially balanced random samples combined into 20 composite samples?

Method: Simulations of 20 independent composites, each containing 5 subsamples, were bootstrapped from the existing data to estimate variance in the mean of 20 composite samples. If the answer to Question No. 2 indicated that variance strata exist, sampling would be specified within these strata. Otherwise, sampling would occur throughout the river without specification of separate strata. The specific steps in the bootstrap approach for a non-stratified design are detailed below.

1. Divide the 5 mi of the LDW into 20 segments of approximately equal area.[9]
2. Subsample within each segment to collect five samples separated by a minimum distance (i.e., the answer to Question 1).
   a. Record the mean for these five samples as the composite sample estimate; treatment of non-detects used substitution at one-half the ~~detection limit (~~DL~~)~~.[10]
   b. Record the standard deviation (SD) for these five samples as the within-composite SD (note: this would not be observed in the baseline sampling, because all individual samples would not be analyzed, although they would be archived).
   c. Record the minimum, maximum, and average distances between samples to verify bootstrap methods.
3. Repeat Step 2 within each of the 20 segments.
4. Store the 20 simulated composites as a single bootstrap replicate of the LDW-wide sample.

[9] These segments were different than the conceptual composite areas proposed for the baseline sampling (Map 3-2 of the main document). The areas on Map 3-2 may not have had enough data points in this dataset to support the bootstrap subsampling (e.g., none of the EAAs were represented in this dataset). The segment boundaries used for this bootstrapping constrained the number of samples available for each random draw (Table A-1). These boundaries were chosen to capture enough data points distributed throughout each segment to ensure that the full range of concentrations within the segment would be represented across the bootstrap replicates. Different segment boundaries could yield slightly different results for any one sample, but the distribution and density of data points in this dataset were large enough that large differences in the overall sampling variance are not expected.

[10] Preliminary simulations compared results between substitution using full and one-half DL to estimate the mean. Due to the high detection frequency, the method used to treat non-detects had very little effect on the outcome.

a. Record the mean, SD, skewness, and kurtosis for the bootstrap sample.

b. Test the goodness-of-fit (GOF) of the bootstrap sample to a normal distribution (Shapiro-Wilk's test), and record the p-value. Non-rejection of the normality test justifies the use of the *t*-interval to estimate the 95UCL for the site-wide mean; otherwise, a 95UCL for a gamma-distributed dataset would be appropriate.

5. Repeat Steps 2 through 4 many times (B = 10,000) to develop a distribution of expected mean and sampling variance.

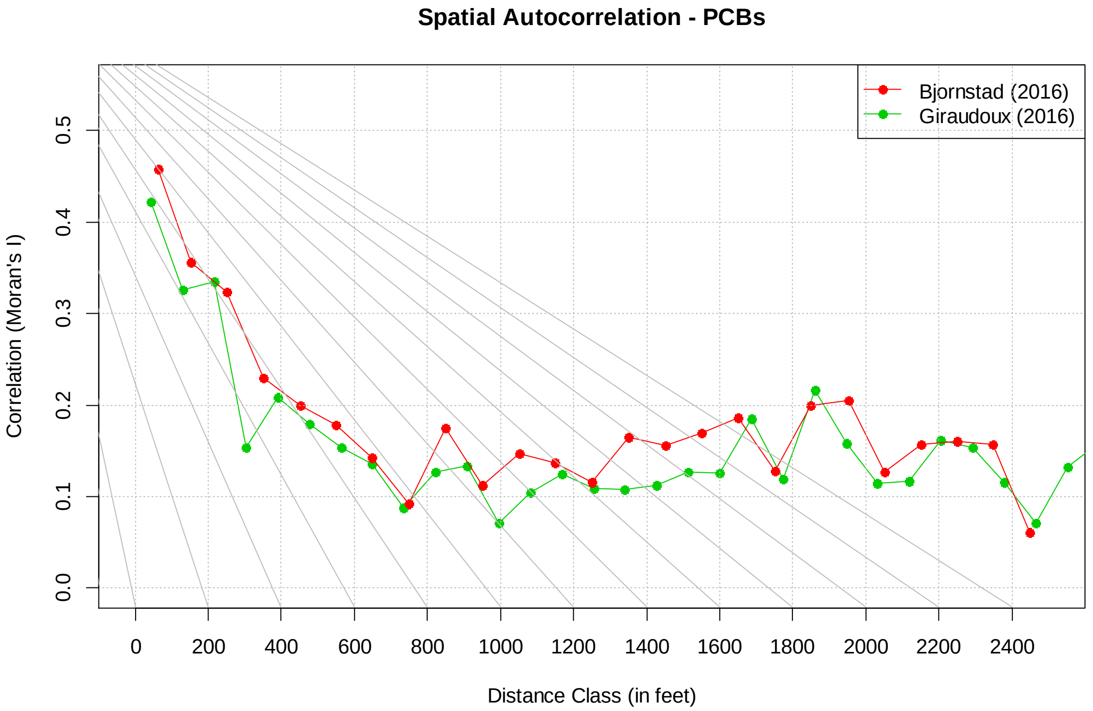
Section ~~4.1~~2.3 presents the results from the analyses described above to answer the preceding questions. The outcome of the GOF test and estimate of the CV for each bootstrap replicate (Step 4) were used to estimate the RME for the mean from a sample design that utilized a spatially balanced collection of 20 composite samples (Section ~~6.1~~2.5).

2.3 Surface Sediments (0–10 cm) Results

The simulation results presented in this section are based on a preliminary ~~used a~~ sampling design of 5 independent samples composited (i.e., averaged) within each of 20 non-overlapping river segments of approximately equal area. The implications of increasing sampling density to increase the number of field samples per composite, the number of analytical composites, or both, are discussed in Section 2.5 where the final sampling design is described.

2.3.1 Question 1

The correlogram for total PCBs in sediments (Figure A-1~~2~~) suggests that the spatial correlation is strongest within approximately the first 200 ft,~~.~~ and that some residual spatial correlation exists at up to 400 -ft. Beyond 4~~3~~00 ft, the correlation ~~is~~ is consistently low (less than 0.20 ~~in the range of 0.15~~) and ~~appears to be~~ within the noise of the random correlations present at ~~further~~ greater distances. Since it appears that samples within 200 ft are, in general, too highly correlated to be considered spatially independent, a minimum separation distance of 200 ft is used for the bootstrap sampling in the subsequent evaluations reported in this appendix. A larger separation may be warranted to ensure independence, but the level of clustering in this dataset is such that using a larger minimum separation distance would severely limit how the sample values could be combined in the simulations.



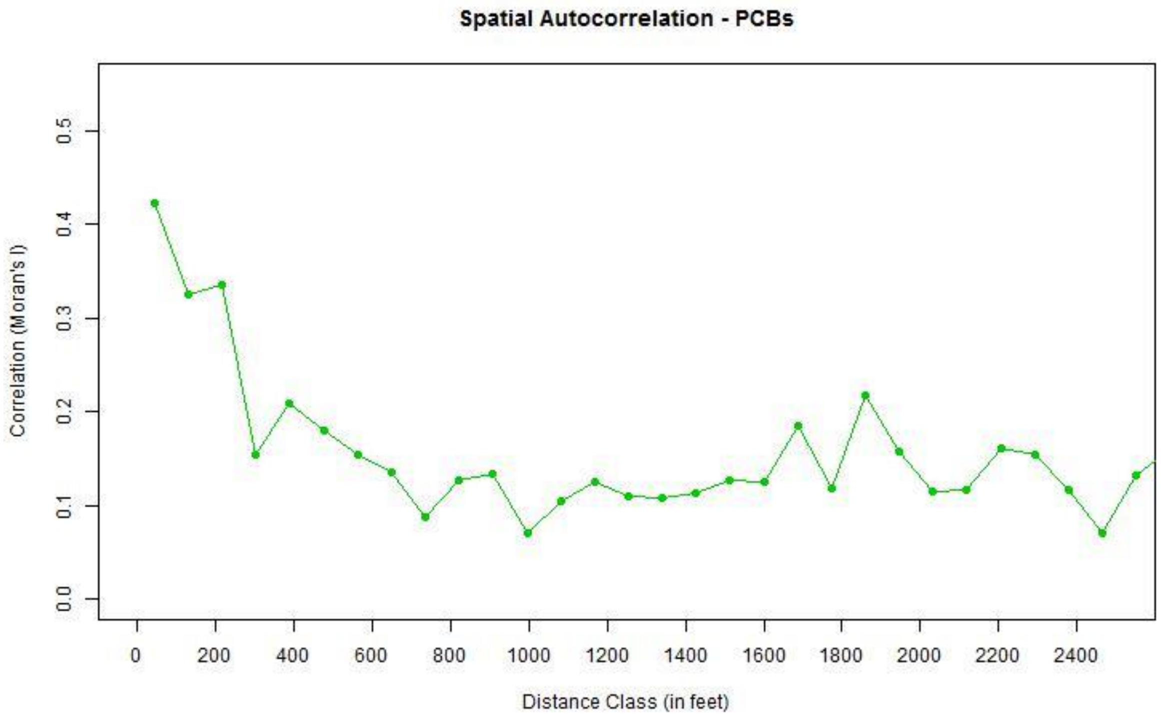


Figure A-1~~2~~. Correlogram (Moran's I versus distance) for total PCBs in the RI dataset

2.3.2 Question 2

The SDs within bootstrapped groups of five samples that were separated by distances between 200 and 1,320 ft (B = 5,000) were plotted against river mile (Figure A-2~~3~~). These results provided a measure of mid-range (200 ft to 0.25 mi)[11] spatial variability across the LDW. This investigation addressed the question of whether variance strata exist within the site. The magnitude of the SDs within sample groups of ~~five~~ 5 was fairly consistent throughout the length of the river (Figure A-2~~3~~). A few exceptions included the areas below RM 0.5 and between RM 2.0 and RM 2.6. These areas with lower variance tended to have fewer samples, so it was assumed that the full variance in these areas was not sampled. These results indicate that dominant variance strata are absent and the entire river can be sampled with the same density throughout.

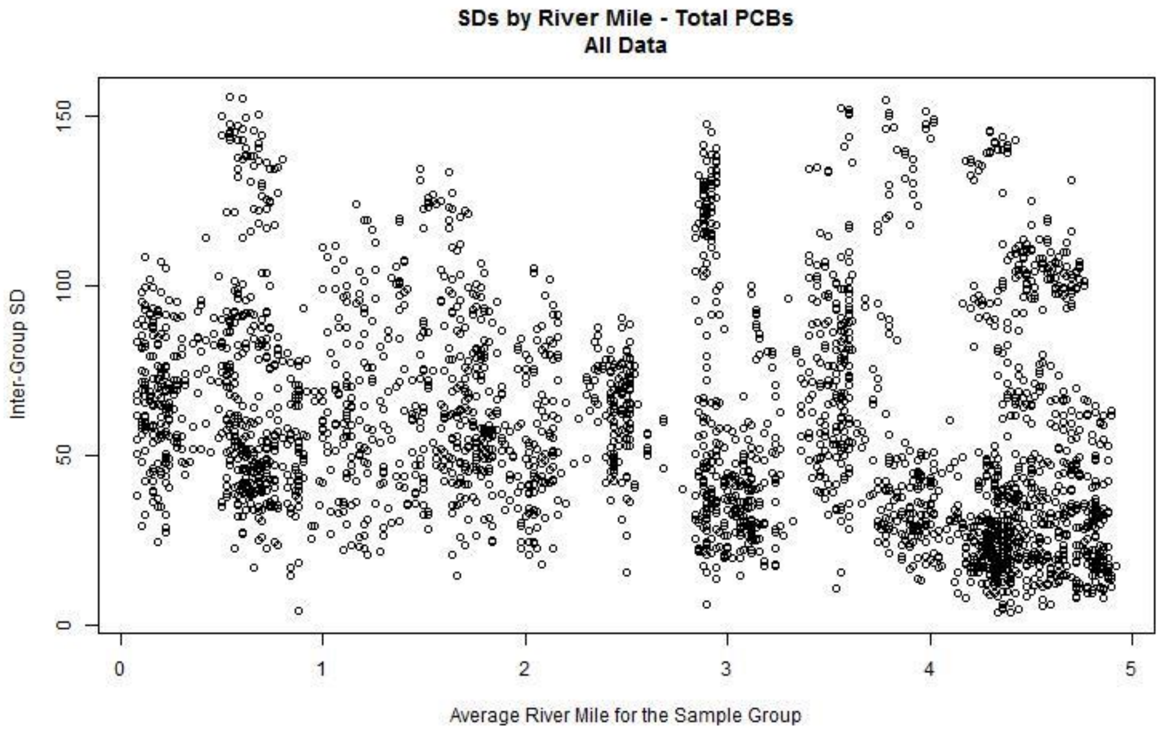


Figure A-2~~3~~. Inter-group SDs for bootstrap sample groups within 200 and 1,320 ft, plotted against the average river mile

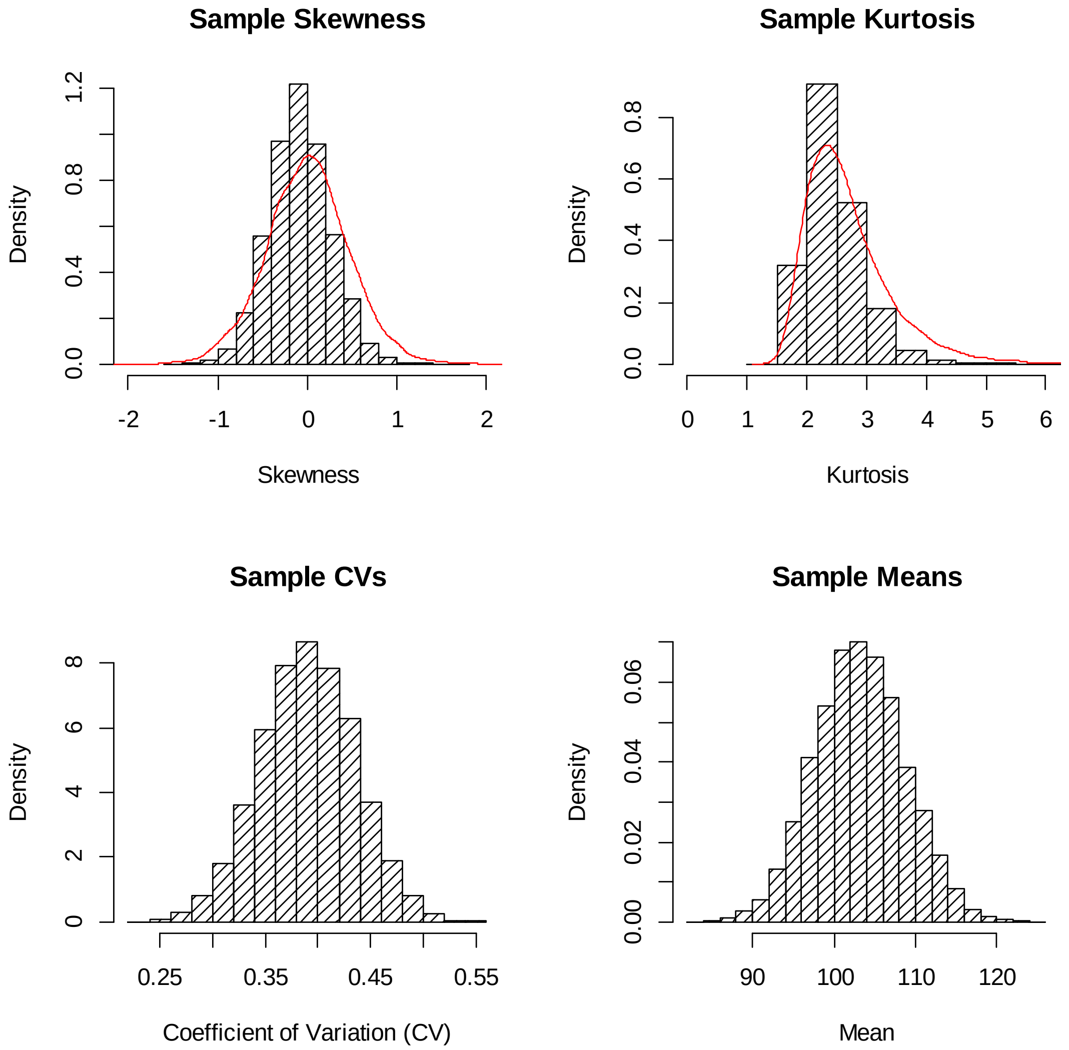
2.3.3 Question 3

Because no strata were identified (via Question 2), LDW-wide bootstrap sampling was conducted ~~as described in Section 3.1~~ using a non-stratified design. The frequency distributions of skewness and kurtosis for each of the bootstrapped samples (size 20)

[11] The approximate scale of separation present among individual samples contributing to a single composite sample in the proposed study design.

indicated that these samples were similar to simulated normal samples of the same size (Figure A-3). The bootstrapped samples were generally symmetric (skewness values near 0, Table A-3) with a tendency for flatter distributions (kurtosis values less than 3) and a low probability of outliers (few kurtosis values greater than 4, Table A-3). The sampling distribution of the mean (Figure A-3) is strongly Gaussian, an expected result based on the Central Limit Theorem (CLT).[12]

[12] The CLT establishes that the mean of a sample randomly drawn (from any distribution) will approach normality as sample size increases.



Sample size of 20 within each bootstrap replicate, B = 10,000. The red line overlaid on the skewness and kurtosis histograms shows values for simulations of normally distributed samples of size 20.

Figure A-3. Frequency distributions of summary statistics (skewness, kurtosis, coefficient of variation, and mean) from each LDW-wide bootstrap replicate

Table A-3. Distribution of skewness, kurtosis, and CV for samples of size 20, across 10,000 bootstrap replicates

	Minimum	1st Quartile	Median	Mean	3rd Quartile	95th Percentile	Maximum
Skewness	-1.4	-0.3	-0.1	-0.1	0.1	0.5	1.7
Kurtosis	1.4	2.1	2.4	2.4	2.7	3.3	6.3
CV	0.23	0.36	0.39	0.39	0.42	0.46	0.55

CV – coefficient of variation

The distribution of the 20 composites within each bootstrap replicate was rejected as normally distributed (Shapiro-Wilks $p < 0.05$) in less than 2% of the bootstrap replicates. This is less than the 5% expected by chance, so these results support the expectation that a set of 20 spatially balanced composite samples from the post-remediated LDW will be a normally distributed sample.

The distribution of sample CVs had an average of 0.4, a 95th percentile of 0.46, and a maximum value of 0.55 across the 10,000 bootstrap replicates (Figure A-3, Table A-3). The average and maximum CVs from this distribution ~~will be~~were used in the sample size estimation presented in Section 2.5~~6~~.

2.4 Surface Sediments (0–10 cm) 95 UCL

Supported by the results in Section ~~4.1.3~~2.3 and the CLT, the sampling distribution of the mean (n = 20 composite samples) is expected to be normally distributed. The *t*-~~-~~interval can be used to calculate the 95UCL of the site-wide mean of a single population as:

$$\mathbf{95}\boldsymbol{UCL} = \bar{\boldsymbol{X}} + \boldsymbol{t}_{(\mathbf{0.05},\boldsymbol{n}-\mathbf{1})}\frac{\boldsymbol{SD}(\boldsymbol{X})}{\sqrt{\boldsymbol{n}}} \qquad \textbf{Equation 1}$$

Where:

$\bar{X}$ is the average of the n samples.

$SD(X)$ is the standard deviation of the n samples.

$t_{(0.05,n-1)}$ is the critical value from the t-distribution with 5% in the upper tail, and n-1 degrees of freedom.

2.5 Surface Sediments (0-10 cm) Discussion

The sampling design for ~~both~~ surface sediment (0–10 cm) is intended to meet the RME target of 25% for the site-wide mean.

The distribution of the mean of 100 samples drawn from the same population is expected to be approximately normal based on the CLT and the law of large numbers. When the 100 samples are combined into 20 averages (composites), the distribution of the mean is still expected to approach normality through the CLT. The bootstrap estimates from existing data that were used to simulate the post-remediated LDW concentration distributions illustrated that the sampling distribution of the mean was indeed Gaussian (Figure A-3). In addition, the 20 composite samples generated in each bootstrap replicate were consistently normally distributed (~~Section 4.1.3~~Section 2.3.3).

The simulations presented in this appendix used existing RI data from the MNR areas. This dataset does not include data from any areas slated for active remedies (i.e., dredging, capping, or enhanced natural recovery [ENR]). So while the MNR dataset used for these simulations is expected to approximate or overestimate the variability post-remediation, it is likely to underestimate the population variance that may be seen during the baseline sampling period. Increasing the sampling density would capture more of this population variability during baseline sampling. The simulations are expected to overestimate the population variance following implementation of the remedy, which will reduce variance in sediment concentrations throughout the LDW since clean sand will be the post-remediation surface in all active remedy areas.

For the stratified random sampling design,[13] the sampling density can be expressed as the range of distances between nearest neighbors. For 100 to 170 grid cells of approximately equal area, the nearest neighbor distances between grid cell centroids were estimated (Table A-4). A desirable sampling density would place samples within 1.5 times the minimum autocorrelation distance, on average. For a minimum autocorrelation distance of 200 ft, 100 grid cells produce a sampling density that averages 1.9 times the minimum separation distance, and ranges from 1 to 3.6 times that distance. For areas with more spatial heterogeneity in the concentrations, this sampling density may be too coarse to capture the variability of concentrations present during baseline sampling. With 140 grid cells, the sampling density increases to an average of 1.4 times the minimum separation distance, and has an approximate range of 1 to 2.6 times that distance. Estimated results for 150, 160, and 170 grid cells are also shown in Table A-4. Based on sampling density considerations, 140 grid cells provide the most cost-effective design for achieving the approximate distance separation targets (within 1.5 times as the minimum separation distance, on average). Note that the actual values may be slightly different from the approximated distance values shown in Table A-4.

Table A-4. Approximate distance between centroids of adjacent grid cells for five different sampling densities

Number of Grid Cells	Minimum Distance (ft)	Maximum Distance (ft)	Mean Distance (ft)	Mean Area per Sample (ac)
100[a]	232 (~1x)[~~a~~b]	726 (~3.6x)	375 (~1.9x)	4.41~~1~~
140	200 (~1x)	520 (~2.6x)	270 (~1.4x)	3.2~~2.9~~
150	200 (~1x)	480 (~2.4x)	250 (~1.3x)	2.9~~2.7~~

[13] Each grid cell is a stratum with a single random sample.

Number of Grid Cells	Minimum Distance (ft)	Maximum Distance (ft)	Mean Distance (ft)	Mean Area per Sample (ac)
160	200 (~1x)	450 (~2.3x)	230 (~1.2x)	2.8~~6~~
170	200 (~1x)	430 (~2.1x)	220 (~1.1x)	2.6

[a] Results for 100 grid cells were calculated based on the preliminary design. Subsequent results in this table were scaled up proportionally and rounded to two significant figures. The shape of the LDW restricts how the grid cells may be arranged to accommodate a target sampling density, so these values are approximations.

[b] Value in parenthesis is the multiplier of the approximate autocorrelation distance of 200 ft that achieves the separation distance shown.

LDW – Lower Duwamish Waterway

A sampling density of 140 grid cells combined into 20 composite samples of 7 samples each was proposed for baseline sampling. This plan provides a sampling interval that randomly varies between 200 and approximately 500 ft, and within approximately 270 ft on average (Table A-4). It avoids severe autocorrelation among the samples (at < 200 ft separation) while capturing the smaller-scale heterogeneity (< approximately 500 ft) for inclusion in each composite sample. Each of the 20 composite samples represents, on average, approximately 22 ac of the site (each with 7 samples). The expected post-remedy RME for the mean using this approach is 18%. Simulations for a sampling approach with 20 composite samples of 7 samples each suggest a lower CV and comparable precision compared to designs using 20 composites of 5 samples each, or 34 composites of 7 samples each (Table A-5).

Table A-5. CV results for simulated data-sets under three different sampling approaches

Total No. Field Samples	No. of Composites	No. of Field Samples per Composite	Median CV	Maximum CV	% with Normality Rejected	% RME Using Maximum CV[a]
100	20	5	0.38	0.53	2%	20%
140	20	7	0.34	0.46	2%	18%
170	34	5	0.45	0.55	7%	17%[b]

[a] Precision estimated using Equation 1.

[b] Uses n = 30.

CV – coefficient of variation

Simulations similar to those detailed in Section 2.2 were conducted to evaluate how the distribution of composite samples would be affected if a greater sampling density was used and the area for each composite sample was reduced. Simulated composite samples from 34 segments,[14] with 5 samples in each composite, resulted in a distribution with greater relative variability than that of the distribution observed for 20 composites of 5 or 7 samples each (CVs shown in Table A-5). The higher normality

[14] The simulation was performed with 34 segments and 170 field samples, because it would have been much more time consuming to assess 30 samples of equal area due of the nature of the dataset. This simulation is provided to illustrate the effect of quantifying spatial variability using a larger number of composites.

rejection rate also indicated the greater tendency for higher skewness in a simulated dataset of 34 samples over that of 20 samples. This skewness is presumably due to more localized conditions being represented by each of the 34 composite samples. The skewness of composite samples from smaller sampling areas would be even more pronounced during baseline sampling, when concentrations in active remedy areas remain elevated, potentially resulting in a more uncertain estimate of the site-wide 95UCL.

The simulation results reported in Section 2.3.3 and Table A-5 support the use of a normal *t*-interval to calculate the 95UCL for the site-wide mean (Equation 1) for 20 composite samples of 5 or 7 samples each. When the data are available and ready to be evaluated, the most appropriate methods to calculate the 95UCL will be determined based on graphical evaluations and GOF tests. For this *a priori* estimation, the 95UCL for the site-wide mean may be expected to be calculated using Equation 1; thus, for ~~the 95UCL and~~ n = 20, the RME is calculated as:

$$\%\boldsymbol{RME} = \boldsymbol{CV}\times\frac{t_{(0.05,20-1)}}{\sqrt{20}}\times\mathbf{100}\qquad \textbf{Equation 2}$$

For CVs ranging from 0.4 to 0.6 (values rounded up from the median and ~~to a value exceeding the~~ maximum CVs observed in the bootstrap results, Table A-5), the ~~proposed~~ sampling design of ~~20 composite samples (~~14~~0~~0 spatially balanced samples, combined into 20 composites of 7~~5~~ samples each, ~~each)~~ is expected to achieve ~~an~~ the targeted 25% RME for the post-remediated site-wide mean (i.e., ~~of approximately~~ 15 to 23%, respectively).

It is important to point out that these results are dependent on the data used for the simulations. EPA has expressed concern that these data may underestimate the variability in PCB concentrations in surface sediment during baseline sampling. The dataset includes only data from MNR areas, which represent approximately 57% of the total LDW (235 of the 412 ac at the site). These areas are away from upland cleanup sites and have lower sediment concentrations suggesting they are not subject the same historical or ongoing sources as areas of the river with higher PCB concentrations, and as such represent the best surrogate available for general ambient variability in the LDW after the cleanup.

Because the MNR dataset excludes data from the active remedy areas (43% of the LDW), it is likely to underestimate the variability expected during baseline sampling (when active remedy areas other than the EAAs still have elevated concentrations). In contrast, however, the MNR dataset is likely to overestimate variability post-remedy, when all active remedy areas have been cleaned up and concentrations are lower. Since the main purpose of this assessment is to estimate variability for long-term monitoring, the sampling design of 20 composite samples with 7 samples each (for a total of 140 samples) was based on ambient variability expected post-remedy, rather than current (baseline) conditions. This design optimized the balance between power, error, and autocorrelation. However, in addition to being focused on the post-remedy condition, it was based on an older dataset that was limited in certain areas, so EPA directed an approach with 24 composite samples with 7 samples each (for a total of 168 samples).

3 Intertidal Sediments (0–45 cm) for Direct Contact During Beach Play and Clamming Scenarios

The intertidal surface sediment (0–45 cm) sampling effort is designed to estimate the 95UCL concentrations for the LDW-wide clamming areas, and the 95UCL concentration for each of the eight beaches. These 95UCL concentrations of human health risk drivers will be used to evaluate cleanup-level compliance for direct contact associated with clamming and beach play RAOs.

Using the compositing plan outlined in Section 3.2.1.4 of the Work Plan (Windward and Integral 2017), the three composite samples will be effectively field replicates of the mean from the sampled locations, either by beach or across all clamming areas. The variance among these composite samples will represent~~s~~ small-scale spatial variability as well as sampling and analytical error, and will be used to calculate 95UCLs at the scale dictated in the ROD.

3.1 Intertidal Sediments (0–45 cm) for Direct Contact During Beach Play 95UCL

Three composite samples (each composite comprised of three, four, five, or nine field samples, depending on beach size) will be available from each beach to estimate the 95UCL for ~~each~~that beach. The shape of the distribution cannot be properly evaluated with only three samples; this level of compositing may be inadequate to invoke the CLT without more information regarding the underlying distributions~~ so the central limit theorem is invoked and normality is assumed~~. Prior to using a 95UCL that assumes a normal distribution, the nature of concentrations from each beach will be investigated and presented in the data evaluation report once the 95UCLs have been calculated. If individual grab samples do not shown high skewness, then it may be appropriate to use a *t*-interval. Otherwise, a non-parametric Chebyshev interval will be used, recognizing that this may result in conservative 95UCLs. This baseline sampling effort will provide an approximate value for the mean at each beach; beach-specific UCLs following the remedy will be more certain as skewness decreases. Information from this baseline sampling may be used to modify future compositing scenarios.

Using this approach, ~~Based on this assumption, ~~the 95UCL will be derived for each beach using either the standard equation for a normally distributed population (Equation 1), or Chebyshev's inequality (Equation 3) with n = 3, and $\bar{X}$ and SD(X) as the mean and ~~standard deviation~~SD, respectively, of the three samples from each beach.

A non-parametric 95UCL for any distribution is provided by Chebyshev's inequality:

$$95UCL = \bar{X} + \sqrt{\left({}^{1}/_{0.05} - 1\right)} \times SE \qquad \text{Equation 3}$$

3.2 Intertidal Sediments (0–45 cm) for Direct Contact During Clamming 95UCL

Three composites samples, each representing the site-wide average, will be used to estimate the 95UCL~~l~~ of the site-wide mean. The shape of the distribution cannot be ~~properly~~ evaluated with only three samples, but these samples (each a composite of 69 field samples) will represent field replicates of the clamming area-wide mean, so the ~~central limit theorem~~CLT ~~is~~ may be invoked and normality ~~is~~ assumed. Based on this assumption, the 95UCL will be derived with the standard equation for a normally distributed population (Equation 1) with n = 3, and $\bar{X}$ and SD(X) as the mean and ~~standard deviation~~SD, respectively, of the three samples across the site.

4 Fish and Crab Tissue

The fish and crab tissue sampling effort is designed to estimate the LDW-wide 95UCL concentrations for comparison to ~~target tissue levels (~~TTLs~~)~~ related to RAO 1 (ROD Table 21[15] (EPA 2014)). The targeted RME for the site-wide mean concentration for fish and crab tissues in the LDW is ≤ 25%, wherein the RME is calculated as the -width of the LDW-wide 95UCL as a percent of the mean.~~[16]~~

To develop the fish and crab tissue sampling design, past data from several LDW tissue sampling efforts (primarily the 2007 RI/FS dataset with additional information for Dungeness crab provided by sample results from 2004 and 2005 (Windward 2010a)) were evaluated. Distributional characteristics of the individual tissue concentrations and site-wide patterns in the mean concentrations were used to identify the best statistical model to identify the sample sizes expected to achieve the targeted RME.

The recommended design includes dividing the LDW into two reaches with four subreaches and creating composite samples of each tissue type within each reach. The reach designations are based on concentration patterns observed in previous tissue data and, where fishing occurs for resident species, per the fishers study (Windward 2016).

Similarly to how baseline sediment data will be used in the future, the site-wide results for baseline tissue sampling will be used to chart, by species, the progression of site-wide tissue concentrations toward the cleanup goals. When sufficient sampling events have been completed (e.g., five or more), the trend for these data can be estimated using regression or correlation methods. In the interim, the baseline data-set may be used most simply in a two-sample, one-tailed comparison to a data-set collected

[15] ROD Table 21 is titled *LDW resident fish and shellfish target tissue concentrations.*

~~[16] See Section 5.2 for more details.~~

in one of the future sampling events. The data should be collected in the same manner over time (i.e., same number of individuals per composite and all same sampling methods) for an “apples to apples” comparison. The specific statistical test used will depend on the nature of the data-sets (e.g., data distribution, equality of variances, number of non-detects) and be appropriate for the stratified sampling design.

4.1 Fish and Crab Tissues Data Used in the Analysis

The 2007 RI fish and crab tissue data were used to assess variability among composites for the tissue types and species targeted in the baseline sampling (Table A-6~~3~~). Data from 2007 were primarily used because earlier data were temporarily elevated following dredging in both the LDW (e.g., Duwamish/Diagonal early action event) and East Waterway. Because of the paucity of information for Dungeness crab in the 2007 dataset, results from 2004 and 2005 (Windward 2010a) were used to provide additional information regarding variance.

In 2007, composite samples were collected within four reaches, with RI reaches T1 and T2 contained within baseline Reach 1 (RM 0.0 to RM 2.9) and RI reaches T3 and T4 contained within baseline Reach 2 (RM 2.9 to RM 5.0). Samples from the different reaches had different mean concentrations, so data from within the RI reaches were appropriately combined using a stratified model to estimate the variability of the site-wide mean for the proposed baseline survey (Table A-6~~3~~). When there are location effects within the population, a stratified model will produce a smaller standard error and hence, a smaller RME. The formulas used to calculate a stratified mean and ~~standard deviation (~~SD~~)~~ are provided in Section ~~5~~4.4. Table A-6~~3~~ provides site-wide estimates of the mean and SD for each tissue type for both stratified models and single stratum models that would be appropriate if there were no differences in mean concentrations among the reaches.

The risk drivers for fish and crab tissues are PCBs and dioxins/furans. In this appendix, results for total PCBs (sum of Aroclors) are the only data evaluated for fish and crab tissues, because the dataset for total PCBs is more robust than the datasets for dioxins/furans.

Table A-6~~3~~. Summary statistics for the 2007 fish and crab tissue total PCB results, including the mean, SD, and CV

Baseline Reach	RI Reach	Tissue Type	N[a]	Mean Concentration (µg/kg, ww)	SD	CV	Comment
Dungeness crab							
1	T1	edible meat	1	15	na	na	4 individuals in this sample
	T1	whole body (calc’d)	1	97	na	na	4 individuals in this sample
	T2		0	na	na	na	no Dungeness caught in this reach

Baseline Reach	RI Reach	Tissue Type	N[a]	Mean Concentration (µg/kg, ww)	SD	CV	Comment
2	T3	edible meat	3	43	6.7	15%	
	T3	whole body (calc'd)	3	234	103	44%	footnote ~~b~~c
	T4		0	na	na	na	no Dungeness caught in this reach
Site wide – single stratum mean and SD		edible meat	4	36	15	42%	
		whole body (calc'd)	4	200	108	54%	
Site wide – stratified mean and SD		edible meat	4	~~36~~29	6.7	23~~19~~ %	~~assumed~~ used SD from T3 for each reach
		whole body (calc'd)	4	~~200~~166	103	6~~5~~2 %	~~assumed~~ used SD from T3 for each reach[c]
		whole body (calc'd)	3	136	21	15%	used SD from T3 with outlier excluded
English sole							
1	T1	fillet with skin	3	343	138	40%	
	T1	whole body	6	525	178	34%	
	T2	fillet with skin	3	293	107	36%	
	T2	whole body	6	693	219	32%	
2	T3	fillet with skin	3	403	78	19%	
	T3	whole body	6	893	364	41%	footnote b
	T4	fillet with skin	0	na	na	na	
	T4	whole body	1	300	na	na	
Site wide – single stratum mean and SD		fillet with skin	9	347	106	31%	
		whole body	19	683	300	44%	
Site wide – stratified mean and SD		fillet with skin	9	361	110	31%	used residual standard error as SD for each reach
		whole body	19	709	266	38%	used residual standard error as SD for each reach
Shiner surfperch							
1a	T1	whole body	6	268	59	22%	
1b	T2	whole body	6	415	115	28%	
2a	T3	whole body	6	763	314	41%	footnote b
2b	T4	whole body	4	315	66	21%	
Site wide – single stratum mean and SD		whole body	22	452	263	58%	
Site wide – stratified mean and SD		whole body	22	440	181	41%	used residual standard error as SD for each reach

Note: Results shaded in blue are the appropriate site-wide estimates for the stratified sampling design that is used in power and sample size calculations.

[a] N = number of composite samples. The numbers of individuals per composite were: 5 individuals (Dungeness crab and English sole) and 10 individuals (shiner surfperch), unless otherwise noted.

[b] High variance was influenced by a single value. Without that value, the CV was greatly reduced, supporting increasing the number of fish per composite in baseline sampling, where feasible.

[c] High variance reported herein was influenced by a single hepatopancreas sample (individual values were 420, 520, and 1020 µg/kg ww). This elevated result is suspect, since this level of variability was not observed in the Dungeness crab composites from 2004 and 2005 datasets. Without that value, the mean and SD were 175 and 21, respectively (CV of 12%).

CV – coefficient of variation
na – not applicable
PCB- polychlorinated biphenyl
RI – remedial investigation
SD – standard deviation
ww – wet weight

4.2 Fish and Crab Tissues Methods

In the baseline sampling to be conducted, English sole and Dungeness crab specimens will be collected and composited within each of two reaches of the LDW: Reach 1 (RM 0.0 to RM 2.9) and Reach 2 (RM 2.9 to RM 5.0) (Map 3-8 of the main document). Shiner surfperch specimens will be collected and composited within each of four subreaches, each comprising one-fourth of the LDW: subreach 1a (RM 0.0 to RM 1.25), subreach 1b (RM 1.25 to RM 2.5), subreach 2a (RM 2.5 to RM 3.75), and subreach 2b (RM 3.75 to RM 5.0) (Map 3-9 of the main document). Four subreaches are being sampled for shiners instead of two because tissue data collected as part of the RI (Windward 2010a) indicated that PCB concentrations and PCB congener patterns showed more spatial differentiation for shiner surfperch than for other fish and crab species analyzed in the RI.

In the 2007 RI/FS dataset, differences in mean concentrations were observed among the reaches (Table A-6~~3~~ and Figure A-4~~1~~). Consequently, a stratified model was the most appropriate model to estimate the site-wide mean and sampling variance for fish and crab tissues, with each reach or subreach having equal weight. A stratified model was applied to the data from the 2007 RI dataset, and means, SDs, and residuals from the stratified model were used to estimate site-wide CVs and examine distributional characteristics of the data (e.g., approximately normal or gamma distributed). Summary statistics (mean, SD, and CV) are summarized in Table A-6~~3~~ by reach, and site-wide estimates are presented for both a stratified model and using a pooled (single stratum) estimate.

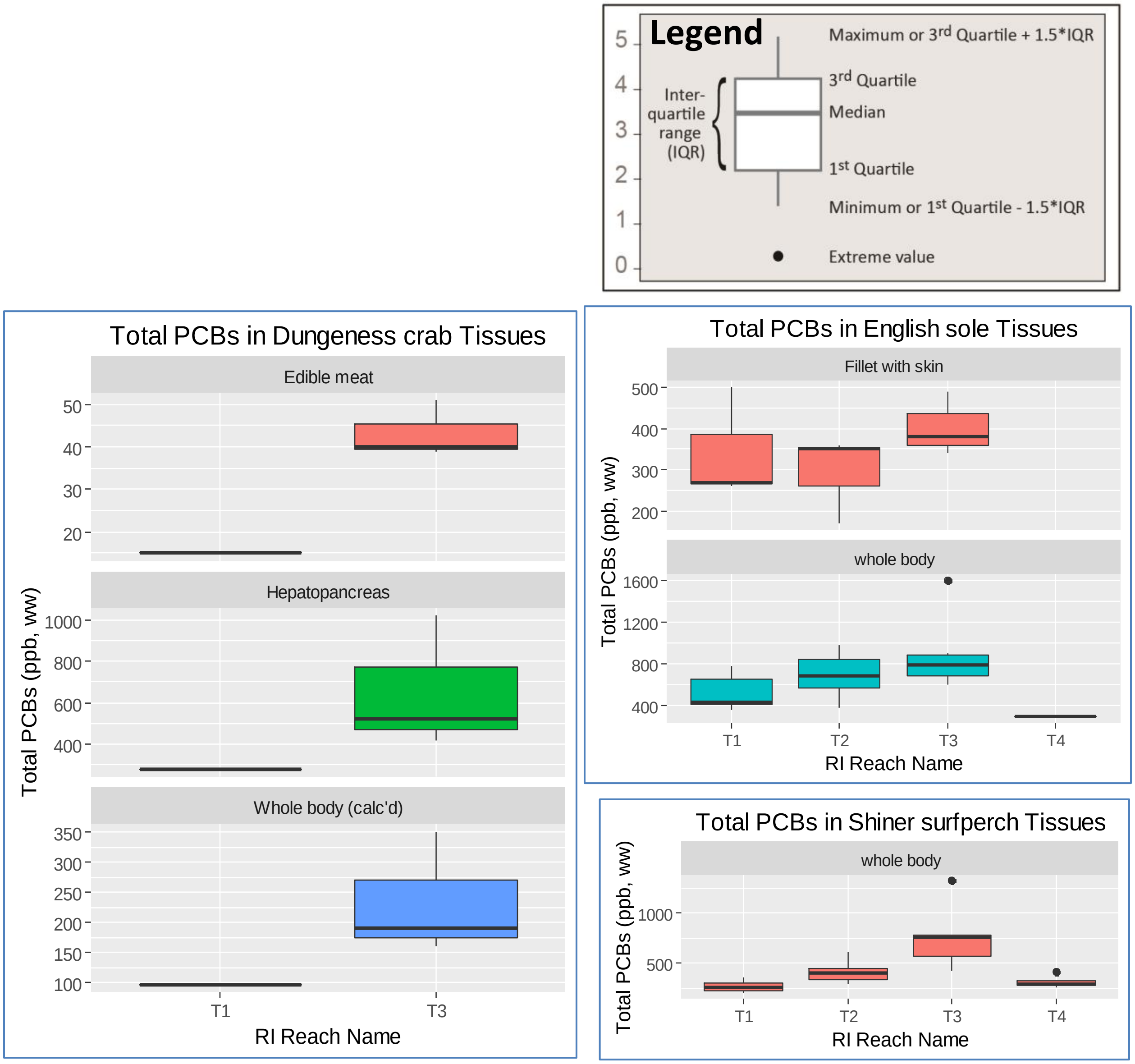


Figure summary statistics are presented in Table A-6~~3~~.

Figure A-4~~1~~. Boxplots showing the distribution by reach of total PCBs (ppb, ww) within each species and tissue type for samples in the RI dataset

A GOF test (Shapiro-Wilk's test for normal distribution) and probability plots (QQ plots) were used to evaluate the distribution of each tissue type for each species. Due to the small sample sizes within each RI reach and evidence that a stratified model was appropriate for the site-wide mean (Table A-6~~3~~ and Figure A-4~~1~~), residuals from the stratified model (the differences between each observation and the reach mean) were combined across all RI reaches to evaluate the statistical distribution for each species and tissue type.

The GOF results are presented in Section ~~4.2~~4.3. The best-fit distribution from the GOF evaluation and estimates of the CV for each tissue type and species were used to generate plots illustrating the expected RME of the mean as a function of sample size (Section 4.5).

4.3 Fish and Crab Tissues Results

Composite tissue concentrations for each species and tissue type appeared to be approximately normally distributed, based on Shapiro-Wilk's GOF test (Table A-7~~5~~) and normal probability plots (Figure A-5~~5~~). Both of these evaluations used the residuals from a stratified model, after excluding two high values identified as outliers (one for English sole, whole body, and one for shiner surfperch, whole body). When the outliers were included, they dominated the probability plots and caused the normality assumption to be rejected. If the tissue data from the baseline sampling effort is skewed, a gamma distribution may be a more appropriate model. Consequently, sample size estimates for both normal and skewed gamma distributions are presented in Section 4.5.

Table A-7~~5~~. Results of the GOF tests on residuals pooled across RI reaches, reported by species and tissue type

Species	Tissue Type	N	Shapiro-Wilk's p-value	Comment
Dungeness crab	edible meat	4	0.31	insufficient data to assess distribution
	whole body (calc'd)	4	0.50	insufficient data to assess distribution
English sole	fillet with skin	9	0.53	data look normal
	whole body	18	0.63	normality rejected for all data; results shown excluding outlier at 1,600 ppb
Shiner surfperch	whole body	21	0.94	normality rejected for all data; results shown excluding outlier at 1,330 ppb

Note: Residuals are the differences between each composite value observation and the mean value of the RI reach.

GOF – goodness-of-fit

ppb – parts per billion

RI – remedial investigation

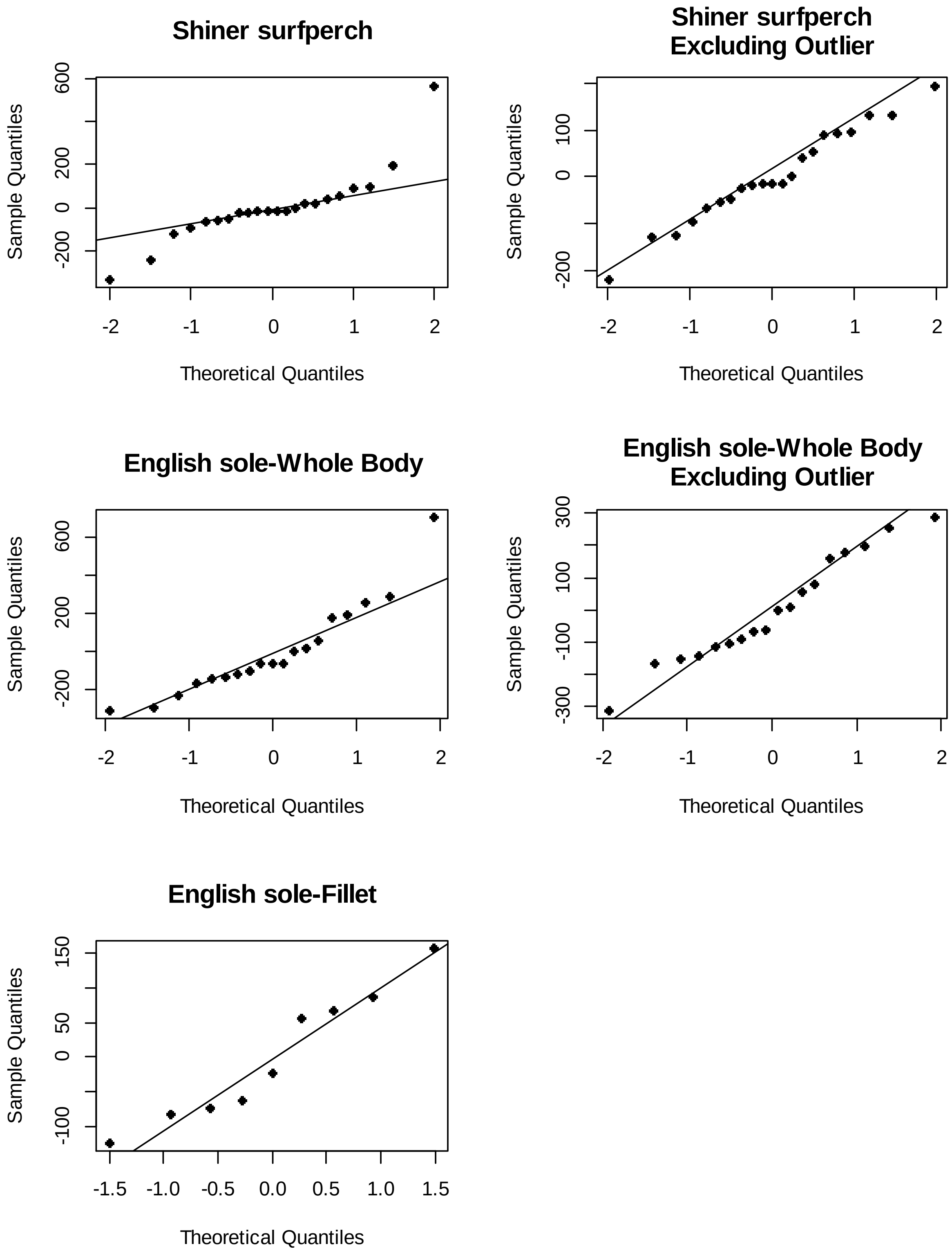


Figure A-5~~5~~. Normal probability plots of the concentration residuals within each RI reach, by species and tissue type for Shiner surfperch and English sole

The compositing methods used in the RI will be modified for the baseline sampling in order to meet the more general objectives of the baseline sampling, and also to reduce variance and the possibility for extreme values. Each composite sample will be comprised of individuals collected throughout the entire reach rather than within smaller subareas. The number of individuals per composite will be increased from 5 to 10 English sole and from 10 to 15 shiner surfperch. More individual crabs per composite sample will not be targeted because of the difficulty in catching the targeted size of Dungeness crab in the LDW.

The changes to the sampling approaches for English sole and shiner surfperch are expected to reduce variance and improve normality from what was observed in the 2007 dataset. The relationship between RME and sample size was calculated and presented for both a normal and a skewed (gamma) distribution (Section 6.2).

The targeted sample size can be identified for each species and tissue type using the curve associated with the appropriate CV value. The applicable CV values derived from the RI dataset were presented in Table A-6~~3~~ and are discussed in more detail below:

- **Dungeness crab – edible meat**: CV ≤ ~~≅~~ 25~~0~~%. There were only three Dungeness crab edible meat composites in the 2007 dataset from which variance could be estimated. These three composites from RI reach T3 had a CV of 15%. Additional information from the 2004 dataset indicated that composite samples from reaches T1 and T3 (n = 3 each) both had CVs of 20%. In 2007, there appeared to be differences in concentrations among reaches, justifying the use of a stratified mean.
- **Dungeness crab – whole body (calculated[17])**: CV < 6~~5~~0%. There were only three Dungeness crab (calculated) whole-body composites in the 2007 dataset from which variance could be estimated. These three composites from R1~~I~~ reach T3 had a CV of 44%, an estimate that was heavily influenced by a single high hepatopancreas result.[18] Additional information from the 2004 and 2005 datasets suggests variability in the calculated whole-body crab values may be much lower than was observed in 2007. The site-wide CV of calculated whole-body values was 12% in 2004 (n = 7), ~~and~~ 4% in 2005 (n = 3), and 15% in 2007 when the outlier was excluded. It appears that there may be much less variability among calculated whole-body crab estimates than suggested by the 2007 results alone, so a CV of ~~50~~60% represents an extreme upper bound, and the actual value is expected to be much lower. Concentration differences were apparent among reaches, lending support to the use of a stratified mean.
- **English sole – fillet with skin**: CV ≅ 30%. Variance was based on three composites from each of three RI reaches (T1, T2, and T3). There did not appear

[17] Each whole-body crab composite concentration was calculated as the weighted sum of separate hepatopancreas and edible meat composites from the same crabs.

[18] The three hepatopancreas results were 420, 520, and 1020 µg/kg wet weight (ww).

to be strong differences in concentrations among reaches. Therefore, if the data support using a single population estimate (instead of a stratified estimated), this approach will gain one additional degree of freedom. Increasing the number of individuals per composite from 5 to 10 should reduce the variability in the baseline survey from what was observed in 2007.

- **English sole – whole body**: CV ≅ 40%. Variance was based on six composites from each of three RI reaches (T1, T2, and T3). Increasing the number of individuals per composite from 5 to 10 should reduce the variability in the baseline survey from what was observed in 2007.
- **Shiner surfperch – whole body**: CV ≅ 40%. Variance was based on six composites from each of three RI reaches (T1, T2, and T3) and four composites from RI reach T4. The mean concentrations within each RI reach were different, and the standard deviations increased with the means, supporting the use of a stratified mean. Increasing the number of individuals per composite from 10 to 15 and compositing throughout each reach should reduce the variability in the baseline survey from what was observed in 2007.

4.4 Fish and Crab tissues 95UCL

The fish and crab tissues will be collected and composited from individual subreaches (shiner surfperch) or reaches (English sole and crab). If it appears that the mean concentrations are different among reaches, stratified estimators will be used to reduce the variance of the site-wide mean.

Using equal weights for each reach, the site-wide mean can be estimated as the grand mean of the mean concentrations within each reach as follows:

$$\boldsymbol{\bar{\bar{X}} = w\sum_{i=1}^{k}\overline{X_\iota}}$$ **Equation 4~~2~~**

Where:

$\overline{X_\iota}$ is the average concentration in reach i (i = 1 to k, where k = 2 for English sole and Dungeness crab; and k= 4 for Shiner surfperch).

$w = 1/k$ (i.e., ½ for sole and crab, and ¼ for perch).

The sampling variance of the stratified mean is:

$$\boldsymbol{\widehat{Var}\left(\bar{\bar{X}}\right) = w^2\sum_{i=1}^{k} S^2_{\bar{X}_i}}$$ **Equation 5~~3~~**

Equation 5~~3~~ simplifies to the following when each of the *k* reaches are weighted equally:

$$\widehat{Var}(\bar{\bar{X}}) = \frac{1}{k^2}\sum_{i=1}^{k} s_{\bar{X}_i}^2$$ **Equation 6~~4~~**

Where:

$s_{\bar{X}_i}^2 = {s_i^2}/{n_i}$

s_i^2 is the usual sample variance estimate of the n_i observations in reach i (i = 1 to k, k = 2 for sole and crab, and k = 4 for perch).

n_i is the sample size in reach i.

For a stratified mean, the CLT is invoked for the UCL estimate (Levy and Lemeshow 1999), although a more conservative Student's *t*-interval is used instead of a Z-interval due to the uncertainty inherent in small samples with an unknown population variance.

$$95UCL = \bar{\bar{X}} + t_{(0.05,df)} \times SE(\bar{\bar{X}})$$ **Equation 7~~5~~**

Where:

$\bar{\bar{X}}$ is the site-wide mean, as calculated above.

SE($\bar{\bar{X}}$) is the standard error of the stratified mean, equal to the square root of the variance estimator in Equation 6~~4~~.

df = the degrees of freedom for this estimator would normally be estimated using Satterthwaite's formula which is a function of variance. For the purposes of this *a priori* sample size estimation, the degrees of freedom will be set to *N* – *k* (*N* = the total number of samples site-wide, *k* = the number of strata).

If the population does not appear to have different means or variances within the different reaches, then the results from all reaches will be pooled for greater power. These pooled data may either be approximately normally distributed (Equation 1), or gamma distributed, which uses the following equations.

$$Approximate\ 95UCL = 2n\hat{k}\bar{X}/\chi^2_{(0.05,df=2n\hat{k})}$$ **Equation 8~~6~~**

Where:

$\hat{k}$ is the shape estimator for the gamma distribution.

$\bar{X}$ is the mean.

$\chi^2_{(0.05,df=2n\hat{k})}$ is the 5th quantile of the chi-square distribution (i.e., 5% of the area is in the left tail), with $2n\hat{k}$ degrees of freedom.

For a gamma distribution, the mean and SD are functions of the scale and shape parameters, ϴ and k, as: $\bar{X} = \Theta k$ and $SD = \Theta\sqrt{k}$. Thus, the CV = $\Theta\sqrt{k}/\Theta k$ = $1/\sqrt{k}$ and k = $1/\mathrm{CV}^2$, and Equation 9~~7~~ expressed in terms of the CV reduces to the following (EPA 2013):

$$\boldsymbol{Approximate\ 95UCL = \frac{2n}{CV^2}\bar{X}/\chi^2_{(0.05,df=2n/CV^2)}}$$

Equation 9~~7~~

And the RME as a proportion of the mean is:

$$\boldsymbol{RME = \frac{(\frac{2n}{CV^2}\bar{X}/\chi^2_{(0.05,df=2n/CV^2)} - \bar{X})}{\bar{X}} = \frac{2n}{CV^2}/\chi^2_{(0.05,df=2n/CV^2)} - 1}$$

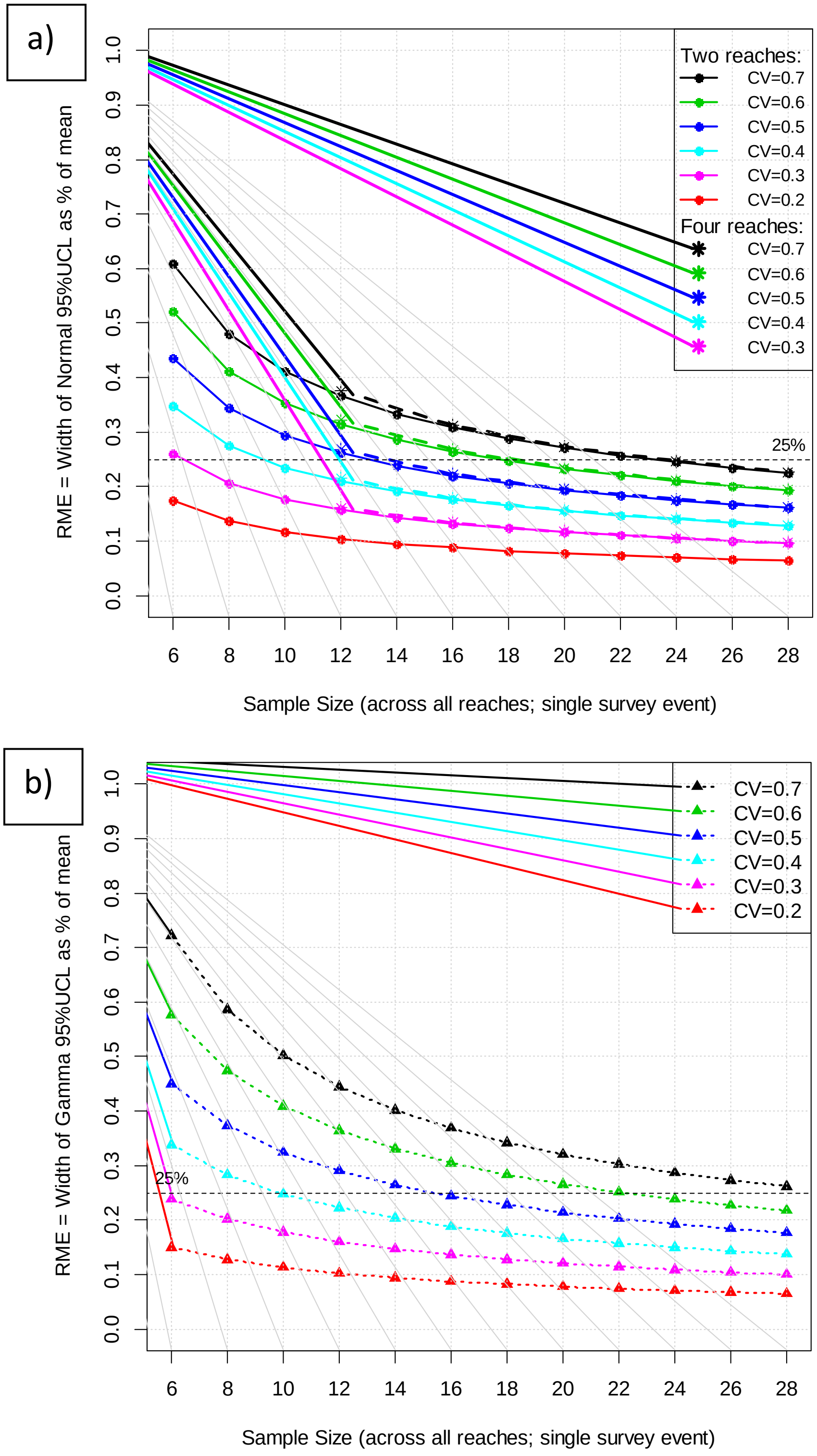
Equation 10~~8~~

4.5 Fish and Crab Discussion

The sampling design for fish and crab tissues is intended to meet the RME target of 25% for the site-wide means, except for whole-body crab, which may have an RME as high as 30%.

The distribution of the fish and crab tissue composites was observed to have outliers in some of the tissue types (Section 4.3). The increase in the number of fish per composite and inclusion of fish across a larger area for each composite is expected to reduce the chance of outliers justifying the use of Student's t-interval for the 95UCL (Equation 7~~5~~). If the baseline data are skewed, the use of a gamma distribution 95UCL will be more appropriate (Equation 9~~7~~). The CVs assumed to be most applicable for these data (Section 4.3) are all less than 0.4~~include the observed extreme values~~, with the exception of the Dungeness crab (calculated) whole-body estimate (CV ≤ 0.6 using all data, and CV < 0.15 excluding the outlier).

Figure A-6 illustrates the relationship between the total number of composite samples and the RME, for normal, stratified estimators of the mean (Figure A-6a) and for a single gamma-distributed population (Figure A-6b), for a range of CVs. Results are displayed for two strata (applicable to English sole and Dungeness crab) and four strata (applicable to shiner surfperch).



Shown for a range of CVs for samples balanced across two or four reaches.

Figure A-6. RME for two or four strata, using a normal UCL (top) and for a single population using a gamma UCL (bottom) versus total sample size

A site-wide total of 12 composite samples of each tissue type for English sole (6 in each of 2 reaches) and shiner surfperch (3 in each of 4 subreaches) is expected to meet the target RME of 25% or better for these species and tissue types, based on CVs of 0.4 or less.[19] The CVs observed in the 2007 dataset were 0.3 to 0.4 for these tissue types, and baseline sampling is expected to be less variable because more individuals will be included in each of the composite samples. If the CV in baseline tissue is greater than anticipated, the analysis of archived tissue, as available, may be recommended (see Section 4.1.2 of the fish and crab QAPP (Windward 2017a)).

A site-wide total of 12 composite samples for Dungeness crab edible meat (6 in each of 2 reaches) ~~is~~ are expected to meet a target RME of approximately 10%, based on a CV of 0.2. The whole-body (calculated) results had high variability in the 2007 dataset (CV of 0.6, Table A-6), but this was influenced by a single high hepatopancreas sample. Information from the 2004 and 2005 datasets suggests that the CV may be much lower (≤ 0.12). All the previous datasets had small sample sizes (n ≤ 3 per reach) due to the difficulty of catching Dungeness crab in the LDW; as a result, the CV estimates are fairly uncertain, ranging from 0.04 in 2004 to 0.62 in 2007, which included an outlier. Based on a CV of 0.6 for the site-wide mean, the RME for whole-body Dungeness crab may be as high as approximately 30%, or less than 10% (based on a CV of 0.15, calculated excluding the outlier).

5 Clam Tissues

The clam tissue sampling effort is designed to estimate the LDW-wide 95UCL concentrations for comparison to TTLs related to RAO 1 (ROD Table 21[20] (EPA 2014)). There will be 1 composite tissue sample[21] for each of the 11 clam collection areas[22] (Map 3-10 of the main document). The site-wide variability for these tissues is currently unknown, so no target RME has been established for these data. The variability observed during baseline sampling can be used to set precision goals for future sampling efforts.

Eleven composite tissue samples will be collected during baseline samplinge, each sample being representative of a single local clam tissue collection area. The 11 samples will be used to calculate the site-wide 95UCL for comparison to target tissue levels. This approach assumes each clam collection area is equally likely to be visited by any person at any point in time over the 30- (non-tribal) and 70-year (adult tribal) exposure periods. Once these data are available, the distribution will be assessed using GOF tests and

[19] Refer to blue shaded rows in Table A-6 for the appropriate CVs for the stratified mean.

[20] ROD Table 21 is titled *LDW resident fish and shellfish target tissue concentrations.*

[21] For arsenic, there will be composite samples of two tissue types (siphon skin and main body minus the siphon skin) from each beach; for all other COCs, there will be composite samples of only one type (whole body), depending on the results of the cPAH clam siphon effort (LDWG 2017).

[22] If clams are not present in clam collection areas within recently remediated areas (i.e., Slip 4, Terminal 117), fewer than 11 areas will be sampled.

probability plots. The 95UCL will be calculated using the most appropriate methods based on the observed distributional characteristics (i.e., distributional form, number of non-detects)~~will be used to calculate the 95UCL~~.

6 Surface Water

The surface water sampling effort is designed, in part, to assess trends in PCB concentrations in surface water. Passive samplers will be deployed at one location in the LDW (RM 3.3).

With limited data available to estimate the variability that the passive samplers will detect in surface water concentrations in the LDW, no target RME was established for this sampling component. Instead, the sampling design was developed using a conceptual model for contaminants in surface waters in the LDW and other available information.

Similar to the surface sediment and tissue sampling efforts, several methods may be used to assess trends for surface water. For example, a simple graphical presentation of surface water concentrations collected over time with an estimate of the slope (or non-parametric correlation) that describes the temporal trend: this would be the simplest way to assess trends, but it would provide only an estimation, and would lack predictions regarding the size of the temporal change or its statistical significance. Another method, which would rely on a statistical test, would compare the mean surface water concentrations from baseline to those in a future sampling period; the sampling design established for this method would provide sufficient statistical power to detect a difference of a meaningful size. The conceptual approach presented herein may be used equally well with either of the described approaches, or others, in a long-term monitoring program. An approximation of the statistical power for this sampling design to detect changes from baseline is provided in Section 6.2 using published data (Apell and Gschwend 2017).

6.1 Surface Water Data Used in the Analysis

As described in more detail in the draft Work Plan (Windward and Integral 2017), the LDW is an estuarine system with a well-stratified salt wedge that is influenced by both freshwater from the Green River upstream and a tidal influx of denser saltwater from Elliott Bay. PCB concentrations in surface water in both the LDW and upstream areas are greater than the lowest applicable or relevant and appropriate requirement (ARAR) identified in the ROD (Tables A-8 and A-9) and are variable (Figure A-7). This variability depends on river conditions, recent precipitation, and the patterns of estuarine circulation.

Table A-8. Summary of ARARs and existing surface water data for PCBs

	PCB Concentration in Surface Water (ng/L)		Notes/Source
	Average	Range	
ARARs:			
WQC – human health[a]	0.064		organism-only and organism + water criteria
WQC – aquatic criteria	30 (chronic)		marine criteria
Washington State aquatic criteria	10,000 (acute); 30 -(chronic)		marine criteria
Upstream (Windward 2017b):			
RM 6.3 – Green River	0.130	0.045–0.514	n = 9; March 2007 to December 2007
RM 10 – Green River[b]	0.618	0.045–6.936	n = 40; September 2011 to February 2015
RM 12.4 – Green River	0.538	0.024–2.434	n = 23; August 2005 to August 2008
LDW (Windward 2010b):			
RM 0.0 – surface (LTKE03)	1.34	0.591–1.947	n = 4; August 2005 to December 2005 (see Table A-2)
RM 0.0 – deep (LTKE03)	0.888	0.250–1.814	n = 3; August 2005 to December 2005 (see Table A-2)
RM 3.3 – surface (LTUM03)	1.14	0.398–1.529	n = 4; August 2005 to December 2005 (see Table A-2)
RM 3.3 – deep (LTUM03)	1.64	0.132–3.211	n = 4; August 2005 to December 2005 (see Table A-2)

[a] ARAR was the most stringent value from the WQC in WAC 173-201(a), NTR, and AWQC at the time of the ROD.

[b] A subset of the samples collected by King County for this RM were biased high due to equipment contamination. These samples were included in the average, but did not impact the range of concentrations presented. Work is ongoing to determine how to correct for this issue (Williston 2017).

ARAR - applicable or relevant and appropriate requirement
AWQC – ambient water quality criteria
LDW – Lower Duwamish Waterway
NTR – National Toxics Rule

PCB - polychlorinated biphenyl
RM – river mile
ROD – Record of Decision
WAC – Washington Administrative Code
WQC – water quality criteria

Table A-9. LDW surface water data for total PCBs (sum of PCB congeners)

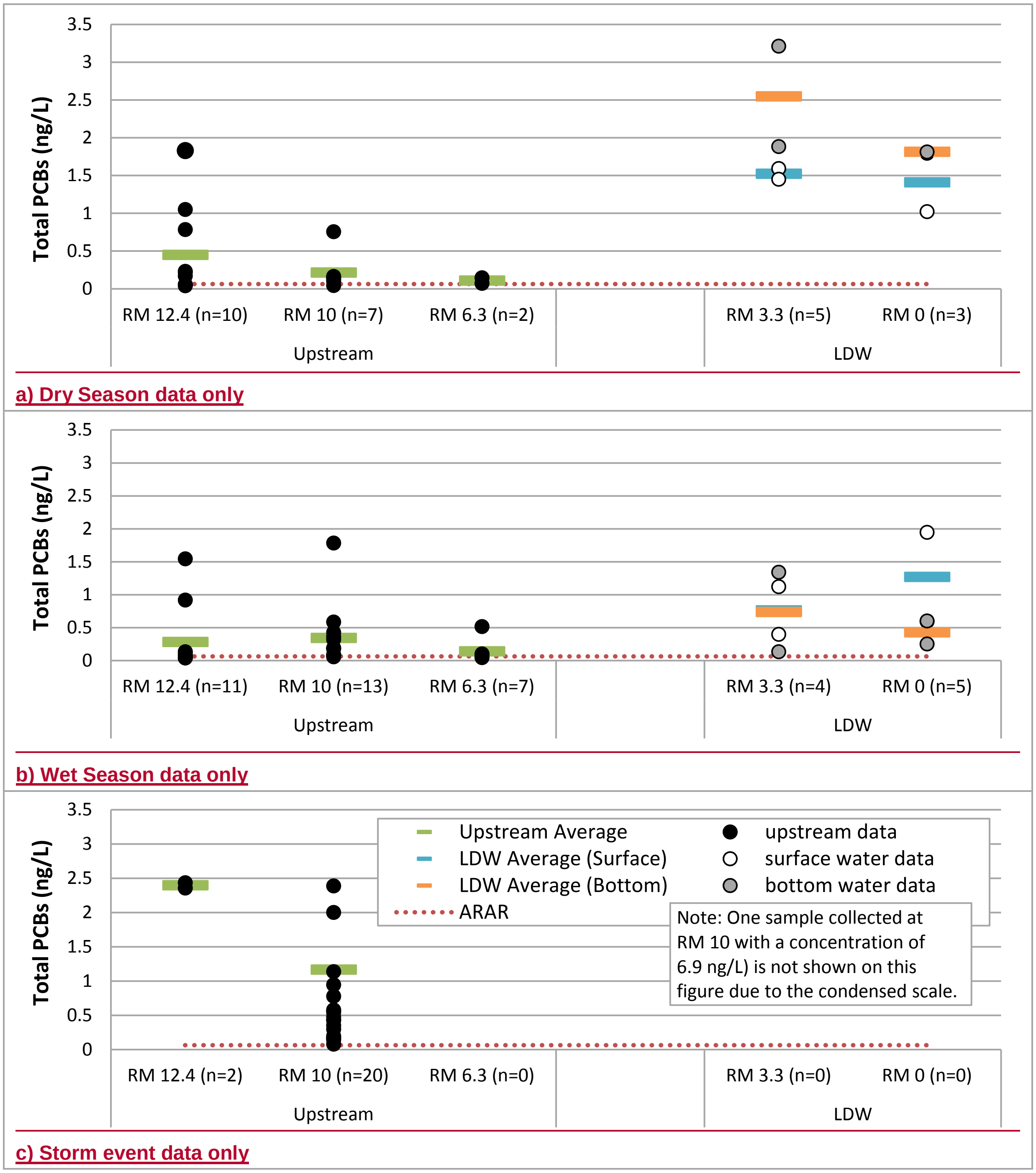
	Total PCB Concentration and Salinity by Date[a]											
	Dry Season Samples						Wet Season Samples					
	8/22/2005 (277 cfs)[b]			9/26/2005 (378 cfs)[b]			11/28/2005 (1,060 cfs)[b]			12/19/2005 (550 cfs)[b]		
Sample Type[c]	Total PCBs (ng/L)	Salinity (PSS)	TSS (mg/L)	Total PCBs (ng/L)	Salinity (PSS)	TSS (mg/L)	Total PCBs (ng/L)	Salinity (PSS)	TSS (mg/L)	Total PCBs (ng/L)	Salinity (PSS)	TSS (mg/L)
LTKE03 (RM 0.0)												
Surface	1.796	22.984	4.8	1.024	25.174	6.0	0.591	13.388	4.2	1.947 J	25.987	5.05
Deep	1.814	28.273	3.1	nc[c]	30.266	3.7	0.25	30.118	2.0	0.599	29.995	2.9
LTUM03 (RM 3.3)												
Surface	1.592 J	16.523	3.4	1.452 J	17.133	5.0	0.398	9.929	4.3	1.122	9.423	4.34
Deep	3.211	26.043	11.1	1.883 J	29.402	5.8	0.132	20.362	4.2	1.341	27.775	3.7

[a] Total PCB concentration represents the sum of detected PCB congener concentrations. RLs for non-detects were not included in the calculation. Data management procedures and data validation criteria were used to calculate the total PCB concentrations presented in the King County technical memorandum (Mickelson and Williston 2006).

[b] Daily mean discharge flow rate in the Green River at USGS Gauge 12113000 in Auburn, Washington.

[c] A number of PCB congener results were rejected because method performance criteria were not met during analysis; therefore, total PCB concentrations were not calculated.

cfs – cubic feet per second
J – estimated concentration
nc – not calculated
PCB – polychlorinated biphenyl
PSS – practical salinity scale
RL – reporting limit
RM – river mile
TSS – total suspended solids
USGS – US Geological Survey



a) Dry Season data only

b) Wet Season data only

c) Storm event data only

Note: One sample collected at RM 10 with a concentration of 6.9 ng/L) is not shown on this figure due to the condensed scale.

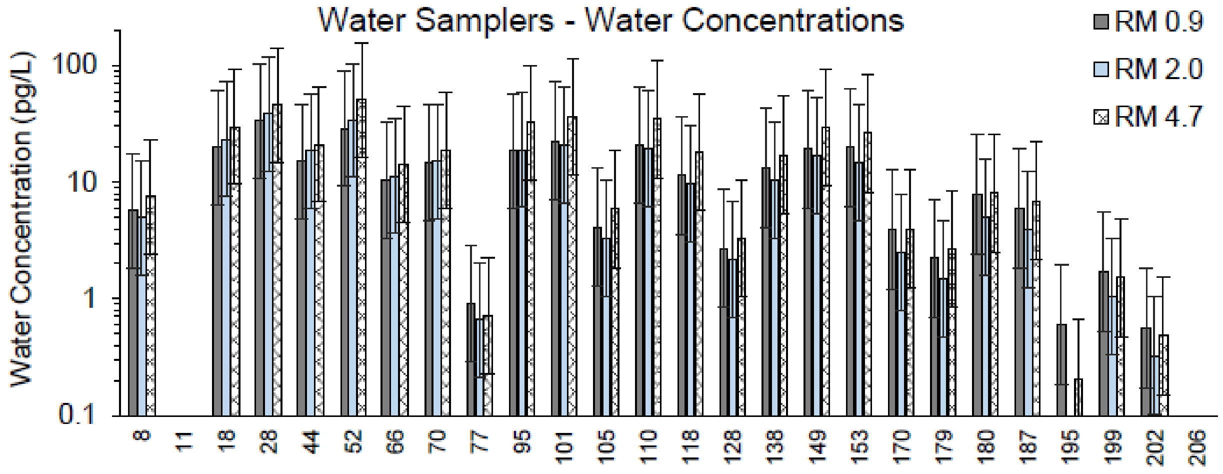
Notes: Storm event data were defined as data from any day with 0.25 in. or more of rainfall. Dry and wet season data were determined based on best professional judgment using information regarding season and rainfall. The ARAR was 0.064 ng/L for human health WQC at the time of the ROD. Surface and bottom water data shown are from the LDW because of the two-layer estuarine flow and greater depth. The upstream data were collected from mid-depth.

Figure A-7. Total PCB concentrations in upstream and LDW surface water samples

The data for PCBs in LDW surface water in Tables A-8 and A-9 and Figure A-7 were from whole-water grab samples. PCB concentrations in whole-water samples are quite variable. Another method to assess PCB concentration trends in surface water is to monitor freely dissolved PCB concentrations using passive sampling devices.

Passive samplers were deployed at three locations (RM 0.9, RM 2.0, and RM 4.7) to evaluate surface water concentrations in the LDW (Apell and Gschwend 2017).

Three replicates samplers were deployed 1 m below the water surface at each location for approximately eight weeks (June 2 to July 27, 2015), after which the samples (referred to as near-surface samples in that study) were analyzed for PCB congeners. As shown in Figure A-8, PCB congener concentrations were similar at the three sampling locations,[23] whereas variability among the concentrations across the three replicates was greater (variability was shown by error bars that represented 95th percentile confidence intervals).



Source: Apell and Gschwend (2017); Figure S4. Error bars represent the 95% confidence intervals for the mean by location.

Figure A-8. Freely dissolved PCB congener concentrations derived from passive samplers deployed in the LDW 1m below the water surface

6.2 SURFACE WATER METHODS

Baseline data for freely dissolved PCB concentrations in surface water will be compared with future long-term monitoring data as follows. An *a priori* power analysis will be used to identify the number of replicate samples expected to provide a reasonable detectable difference for this comparison. *A priori* power analyses are predictive, describing a scenario for an expected result with a given level of confidence; the accuracy of this prediction in a particular situation is dependent on whether the assumptions that were made about mean and variance are valid. The estimated

[23] The uncertainty analysis included an assessment of uncertainties associated with analytical measurements and partition coefficients.

replicate sample size identified through the power analysis will rely on limited existing information about the variability expected among field replicates in passive samplers in the LDW.

The baseline mean concentrations will be estimated as the average of replicates over two dry-season passive sampler deployments; the future mean concentrations (i.e., post-remedy) will be estimated as the average of replicates over two dry-season deployments at the same location. For example, if baseline data are collected in August 2017 and August 2018, these data will be compared with future data collected over two consecutive dry-season passive sampler deployments.[24]

The data generated during baseline characterization and any future sampling period may be compared using a parametric *t*-interval for an equation that estimates the difference between the means of two time periods (i.e., a two-tailed, two-sample comparison, similar to a simple *t*-test but modified to use estimates of mean and standard error [SE] that are appropriate for the sampling design and difference equation being tested). This comparison between the future and baseline summer means for a single station and depth is a two-tailed hypothesis test that has the following null and alternative hypotheses:

$$H_0: \mu_{\text{future}} = \mu_{\text{baseline}}$$

Vs.

$$H_a: \mu_{\text{future}} < \mu_{\text{baseline}} \text{ or } \mu_{\text{future}} > \mu_{\text{baseline}}$$

When the grand mean (a mean of two annual means) from baseline sampling is compared to the grand mean from a future timeframe, the difference equation (Δ) can be written as:

$$\Delta = \frac{1}{2}(\bar{\mathbf{S}}_{\mathbf{B1}} + \bar{\mathbf{S}}_{\mathbf{B2}}) - \frac{1}{2}(\bar{\mathbf{S}}_{\mathbf{F1}} + \bar{\mathbf{S}}_{\mathbf{F2}}) \qquad \textbf{Equation 11}$$

Where:

$\bar{S}_{Bj}$ = mean for a given station and depth during baseline year j (j=1 or 2)

$\bar{S}_{Fj}$ = mean for the same station and depth during future year j (j=1 or 2).

Replication occurs within the station, depth, and year, such that the variability among field replicates within a station is the scale against which the difference in means (Equation 11) is evaluated. Using the relationship that the variance of a sum is the sum of the variances for independent samples, the SE of this difference equation is estimated as:

[24] The need for data from two consecutive dry seasons will be evaluated over time.

$$\boldsymbol{\widehat{SE}(\Delta) = \sqrt{\textstyle\sum_j c_j^2 S_j^2/n_j}}$$ Equation 12

Where:

c_j = coefficient for the j^{th} mean in the difference equation (Equation 11), either ½ or -½

S_j^2 = variance among field replicates for the j^{th} sampling period; if variances are equal, a single pooled residual variance estimate, S_p^2, can be used for each group

n_j = number of field replicates within the j^{th} sampling period; replication is designed to be equal within every sampling period and location, but sample sizes may be unequal in the final analysis if samplers are lost

To establish the number of samples needed to provide an expected minimum detectable difference (MDD) between a baseline mean and a future mean, the following relationship is used:

$$\boldsymbol{MDD \geq \widehat{SE}(\Delta)\left(t_{\alpha(2),df} + t_{\beta(1),df}\right)}$$ Equation 13

Assuming equal variances and equal n during all sampling periods, this simplifies to:

$$\boldsymbol{MDD/S_p \geq \left(t_{\alpha(2),df} + t_{\beta(1),df}\right)/\sqrt{n}}$$ Equation 13a

This is the scaled MDD (i.e., the MDD expressed in units of the square root of the pooled residual variance), where:

df = the degrees of freedom associated with the standard error estimate (Equation 12)

Types I (α) and II (β) errors = 10%

When the scaled MDD is multiplied by the baseline coefficient of variation (CV = SD/mean), and it is assumed that the baseline SD is similar to the pooled residual SD (S_p), then the MDD is expressed as a percentage of the baseline mean:

$$\boldsymbol{MDD/S_p \times S_p/Mean = (Mean_{Baseline} - Mean_{Future})/Mean_{Baseline}}$$

$$\boldsymbol{\geq CV \times \left(t_{\alpha(2),df} + t_{\beta(1),df}\right)/\sqrt{n}}$$ Equation 13b

If the data must be log-transformed to meet the normality assumption for the residuals, then the MDD is the minimum difference between the mean of the log-scale values that would be detected with the specified Type I and II error rates. Exponentiation of the log-scale MDD (MDD′) yields:

$$\boldsymbol{exp(MDD') = exp\left(Mean_{log(Baseline)} - Mean_{log(Future)}\right)}$$

$$= \boldsymbol{GeoMean_{Baseline}/GeoMean_{Future}} \qquad \textbf{Equation 14}$$

And

$$(\boldsymbol{GeoMean_{Baseline} - GeoMean_{Future}})/\boldsymbol{GeoMean_{Baseline}}$$

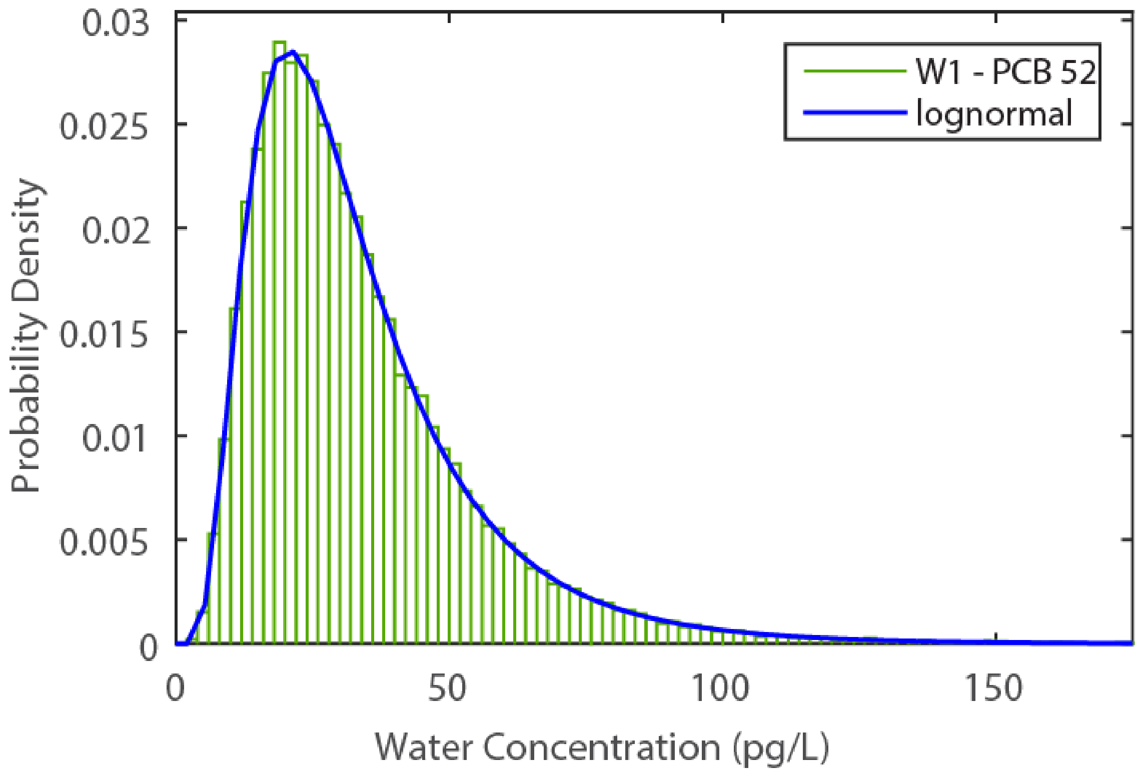
$$= \mathbf{1} - \mathbf{1}/\boldsymbol{exp}(\boldsymbol{MDD'}) \qquad \textbf{Equation 14a}$$

Where:

$GeoMean_p$ = the geometric mean for period p~~.~~

Hence, MDD' for log-transformed data is computed using Equation 13a, the result of which is then converted to a percent difference of geometric means on the original scale using Equation 14a.

The estimated water concentrations from passive samplers are likely to be approximately left skewed (log-normal) for some individual congeners, due to log-normal errors in the estimated partition coefficients that are used to estimate the water concentrations. Figure A-9 (Figure S6 from Apell and Gschwend (2017)) shows the results for a single PCB congener with simulated analytical errors. The estimate of total PCBs is a sum of congeners, which may also be left skewed. Power results are presented assuming both normal and log-normal distributions for the data.



Source: Figure S6 in Apell and Gschwend (2017)

Figure A-9. Histogram and fit of PCB-52 water concentrations measured with passive samplers when the error is propagated with a randomized simulation

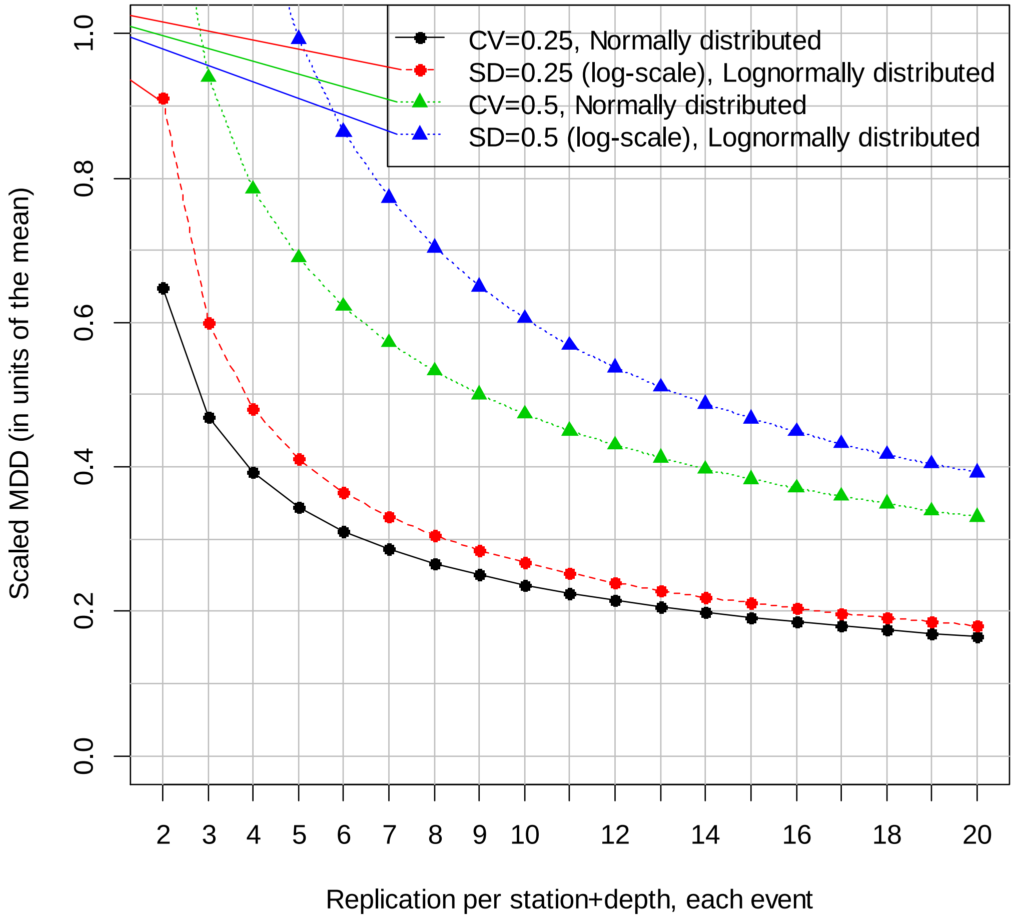
Analytical results that are approximately normally distributed may be compared using a *t*-interval, and the relationship between sample size and MDD as a percent of the baseline mean is described by Equation 13b. On the other hand, if the water concentration results are log-normally distributed, then the comparison would use a *t*-interval for the log-transformed data, and the relationship between sample size and MDD of the geometric means as a percent of the baseline geometric mean would use the relationship shown in Equation 14a.

6.3 Surface Water Results

In the LDW, the CV for total PCBs measured in three passive samplers to be placed at approximately RM 0.9, RM 2.0, and RM 4.7 was inferred to be on the order of 25% (based on the results provided in Apell and Gschwend (2017)[25]). Figure A-10 shows the

[25] In Apell and Gschwend (2017), total PCBs appear to be a sum of 27 PCB congeners. RKM 1.4, RKM 3.2, and RKM 7.6 reported in the document are equivalent to RM 0.9, RM 2.0, and RM 4.7, respectively, and the CV was approximated for the three samples based on a reported range from 0.28 to 0.42 ng/L and a

MDD as a percent of the baseline mean for both a normal and log-normal assumption regarding the data distribution. For a normal distribution, the MDD expressed as a percent of the baseline mean assumes a CV of 25% for field replicates; for a log-normal distribution, the MDD assumes a log-scale SD of 0.25.[26] Additional curves are shown for a CV of 50% and a log-scale SD of 0.5 to reflect the possibility that field variability is much higher than that expressed by the limited data that is currently available.



Note: Assumes a parametric *t*-interval test for the difference of means between baseline (2 years) and future (2 years) for data that are either normally or log-normally distributed. Types I and II errors are both set at 10%. The CV and log-scale SD values of 0.25 are comparable to reported field variability.

Figure A-10. Relationship between replication within each station/depth and sampling event versus scaled MDD (expressed in units of the mean)

With a balanced design (2 years in baseline and 2 years in the future), 9 field replicates from each sampling event (for a total of 18 results during the 2 baseline years, and 18 results during the 2 future years) are expected to result in an MDD equivalent to

geometric mean of 0.32 ng/L. The middle result was estimated as 0.28 ng/L; so SD (0.28, 0.28, 0.42)/mean (0.28, 0.28, 0.42) = 25%.

[26] SD(log(0.28), log(0.28), log(0.42)) = 0.23, which is rounded up to 0.25.

approximately 25% of the baseline mean if the CV = 0.25, and 50% of the mean if the CV = 0.5, for normally distributed data. If the data are log-normally distributed, the predicted MDD is higher, ranging from 28 to 65% of the baseline geometric mean (for log-scale SDs of 0.25 and 0.5, respectively).

Assuming a mean (or geometric mean) baseline value of approximately 0.32 ng/L for total PCBs in the LDW (Apell and Gschwend 2017), nine field replicates from one station (for both each of 2 years in baseline and each of 2 years in the future) are expected to result in a detected a minimum difference of approximately 0.1 ng/L (using field variability reported by Appel and Gschwend, and either a normal or log-normal distribution).

7 References

Apell JN, Gschwend PM. 2017. The atmosphere as a source/sink of polychlorinated biphenyls to/from the Lower Duwamish Waterway Superfund site. Environ Pollut 227:263-270.

Bjornstad ON. 2016. ncf: spatial nonparametric covariance functions. R package version 1.1-7 [online]. Updated April 13, 2016. Available from: https://cran.r-project.org/web/packages/ncf/index.html.

EPA. 2013. ProUCL Version 5.0.00. Statistical software for environmental applications for data sets with and without nondetect observations. EPA/600/R-07/041 [online]. Office of Research and Development, US Environmental Protection Agency, Washington, DC. Updated September 2013. Available from: http://www.epa.gov/osp/hstl/tsc/software.htm.

EPA. 2014. Record of Decision. Lower Duwamish Waterway Superfund Site. US Environmental Protection Agency.

EPA, SERDP, ESTCP. 2017. Laboratory, field, and analytical procedures for using passive sampling in the evaluation of contaminated sediments: user's manual. EPA/600/R- 16/357. February 2017 final web version (1.0). US Environmental Protection Agency, US Department of Defense, Strategic Environmental Research and Development Program, and Environmental Security Technology Certification Program.

Ghosh U, Driscoll SK, Burgess RM, Jonker MTO, Reible D, Gobas F, Choi Y, Apitz SE, Maruya KA, Gala WR, Mortimer M, Beegan C. 2014. Passive sampling methods for contaminated sediments: practical guidance for selection, calibration, and implementation. Integr Environ Assess Manag 10(2):210-223.

Giraudoux P. 2016. pgirmess: data analysis in ecology. R package version 1.6.5 [online]. Updated September 25, 2016. Available from: https://cran.r-project.org/web/packages/pgirmess/index.html.

Gschwend P, Adams EE, Michalsen M, von Stackelberg K. 2016. Combining mass balance modeling with passive sampling at contaminated sediment sites to evaluate PCB sources and food web exposures. Stragegic Environmental Research and Development Program.

King County. 1999. King County combined sewer overflow water quality assessment for the Duwamish River and Elliott Bay. Vol 1: Overview and interpretation, plus appendices. King County Department of Natural Resources, Seattle, WA.

LDWG. 2017. Clam sampling for cPAH analysis of siphon skin. Lower Duwamish Waterway Group, Seattle, WA.

Levy PS, Lemeshow S. 1999. Sampling of Populations: Methods and Applications. 3rd ed. Wiley Series in Survey Methodology. Wiley.

Mickelson S, Williston D. 2006. Technical memorandum: Duwamish River/Elliott Bay/Green River water column PCB congener survey: transmittal of data and quality assurance documentation. King County Department of Natural Resources, Seattle, WA.

R Core Team. 2016. R: A language and environment for statistical computing [online]. R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org/.

Williston D. 2017. Personal communication (phone conversation between D. Williston, King County, and S. Replinger, Windward, regarding King County water samples collected in the Green River that were biased high due to equipment contamination). Windward Environmental LLC and King County Water and Land Resources Division/DNRP, Seattle, WA. May 15, 2017.

Windward. 2010a. Lower Duwamish Waterway remedial investigation. Remedial investigation report. Final. Prepared for Lower Duwamish Waterway Group. Windward Environmental LLC, Seattle, WA.

Windward. 2010b. Lower Duwamish Waterway remedial investigation. Remedial investigation report. Final. Prepared for Lower Duwamish Waterway Group. Appendix I. Source control area-related facility information. Windward Environmental LLC, Seattle, WA.

Windward. 2016. Lower Duwamish Waterway fishers study data report. Final. Windward Environmental LLC, Seattle, WA.

Windward. 2017a. Baseline fish and crab tissue collection and chemical analyses - quality assurance project plan. Draft. Submitted to EPA on May 12, 2017. Windward Environmental LLC, Seattle, WA.

Windward, Integral. 2017. Pre-design studies work plan. Lower Duwamish Waterway Superfund site. Draft. Prepared for the Lower Duwamish Waterway Group for submittal to EPA Region 10. Windward Environmental LLC and Integral Consulting Inc., Seattle, WA.

Windward. 2017b. Technical memorandum: compilation of existing data. Draft final. Windward Environmental LLC, Seattle, WA.